**Heart Attack Risk Factors Analysis**

Bowen Duan, Malak Parmar, Yang Wu

Group 2

Applied Machine Intelligence, College of Professional Studies

Northeastern University

ALY6040 [70413 Data Mining Applications](https://northeastern.instructure.com/courses/160699)

Kasun Samarasinghe

Nov 5, 2023

**Introduction**

This report presents a comprehensive EDA of a dataset containing health-related metrics potentially associated with heart attack risk. We aim to predict the risk of heart attack based on various health indicators and personal habits and build models to help identify individuals at higher risk, allowing for preventive measures or further medical examination.

The dataset comprises various health indicators and lifestyle factors of 8763 individuals. The data was sourced from Kaggle and is believed to be useful for healthcare providers to predict and prevent heart attacks.

Each feature across all records has a valid data entry, which indicates that the dataset is well-prepared and is clean for analysis. We also verified that the dataset has no duplicate records since duplicate records can skew the analysis and potentially lead to incorrect conclusions. The uniqueness of each data in the data set ensures that our analysis is based on the actual variance of the data.

Each of the 8763 patient records in our dataset contains 26 features. Those features cover demographic, medical, and lifestyle information. This volume is substantial enough to conduct analysis and find meaningful insights.

Now we will delve into the specifics of the data attributes to provide you with a comprehensive understanding of the dataset. The following data dictionary table outlines each feature present in our dataset, it should help us understand the variables and their respective roles in our analysis.

**Data Cleaning**

**Data dictionary**

Before cleaning the data, we went through all the features and clarified the meaning of the data type. Please refer to below Table 1.

**Table 1**

Description and data type for each column in the dataset

| Feature Name | Description | Data Type |
| --- | --- | --- |
| Patient ID | A unique identifier is assigned to each patient. | String |
| Age | The patient's age is in years. | Integer |
| Cholesterol | The patient's cholesterol level is measured in milligrams per deciliter (mg/dL). | Integer |
| Blood Pressure | The patient's blood pressure reading is represented as systolic over diastolic (mmHg). | String |
| Heart Rate | The number of heartbeats per minute. | Integer |
| Diabetes | Indicator of whether the patient has diabetes (0: No, 1: Yes). | Boolean |
| Family History | Indicator of whether there is a family history of heart disease (0:No, 1: Yes). | Boolean |
| Smoking | Indicator of whether the patient smokes (0: No, 1: Yes). | Boolean |
| Obesity | Indicator of whether the patient is considered obese based on BMI (0: No, 1: Yes). | Boolean |
| Alcohol Consumption | Indicator of whether the patient consumes alcohol (0: No, 1: Yes). | Boolean |
| Exercise Hours Per Week | The average number of hours the patient exercises each week. | Float |
| Diet | The general categorization of the patient's diet (e.g., Healthy, Unhealthy, Average). | String |
| Previous Heart Problems | Indicator of whether the patient has had previous heart problems (0: No, 1: Yes). | Boolean |
| Medication Use | Indicator of whether the patient is on medication (0: No, 1: Yes). | Boolean |
| Stress Level | A numeric score representing the patient's level of stress. | Integer |
| Sedentary Hours Per Day | The average number of hours the patient is sedentary each day. | Float |
| Income | The patient's income level or category. | Integer |
| BMI | The patient's Body Mass Index. | Float |
| Triglycerides | The level of triglycerides in the patient's blood. | Integer |
| Physical Activity Days Per Week | The number of days per week the patient engages in physical activity. | Integer |
| Sleep Hours Per Day | The average number of hours the patient sleeps each day. | Integer |
| Country | The country where the patient resides. | String |
| Continent | The continent where the patient's country is located. | String |
| Hemisphere | The hemisphere (Northern/Southern) is based on the patient's location. | String |
| Heart Attack Risk | The assessed risk of heart attack for the patient (0: Low, 1: High). | Boolean |

**Handle Missing Values**

After a thorough assessment of the dataset, it is revealed that it has 8763 entries. We found the dataset to be fairly clean, with no missing values by using:

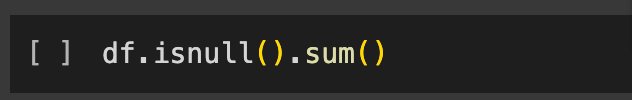


Fig.1 Null cell check

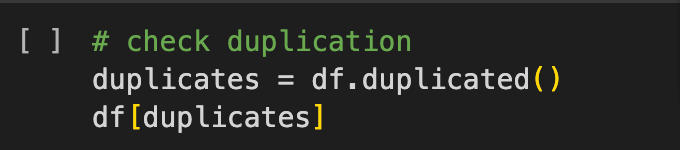


Fig.2 Duplication check

For text-based anomalies, we have features like “Country”, “Continent”, and “Hemisphere”, so we check the uniqueness of the content for them to see if any misspellings, case sensitivity, or space errors. For example

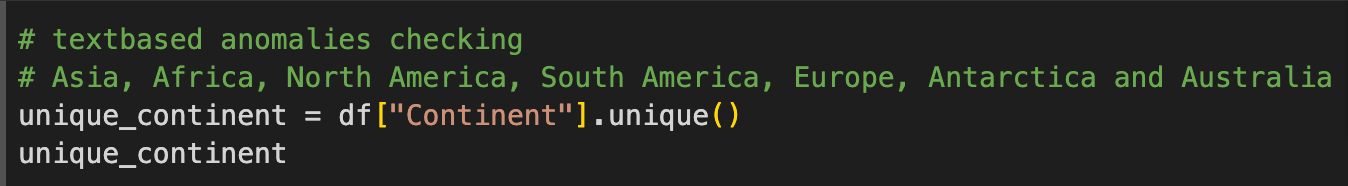


Fig.3 Data Cleaning in column Continent

However, for the column ‘Continent’, the spelling format is found to be inconsistent, for example, we noticed some entries spelled ‘North America’ as ‘nortH America’. To deal with this, we wrote functions to format all of the entries to be in the title case.

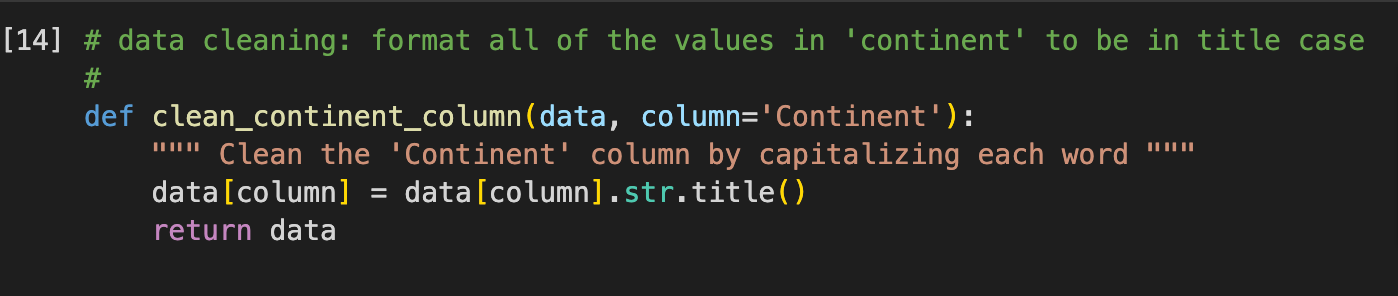


Fig.4 Data Cleaning in column Continent

Other than that, all of the entries are correctly formatted, and the data types are consistent with the expected formats: numerical values for quantifiable health metrics, boolean values for binary categorical variables, and object data types for nominal features.

**Outliers and Suspicious Data**

We employed both visual and statistical methods to identify outliers across various numerical variables. Initially, the box plots suggested the presence of potential outliers in several columns including ‘Cholesterol’, ‘BMI’, ‘Triglycerides’, ‘Income’, and ‘Sedentary Hours Per day’. These data points appeared to be distant from the bulk of the distribution. However, we decided to apply the Interquartile Range method to be more precise, and it showed no outliers. We believe the discrepancy suggests that while the box plots visually indicate extreme values, they actually did not meet the statistical threshold of 1.5 times the IQR beyond the quartiles to be classified as outliers.

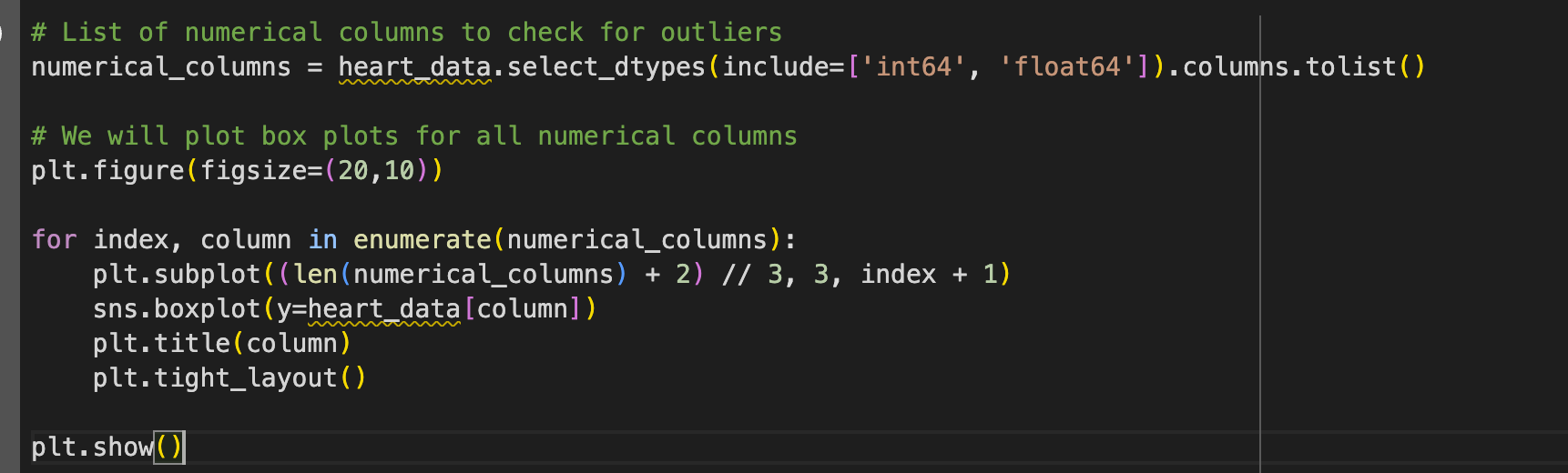


Fig.5 Outlier check for numerical feature

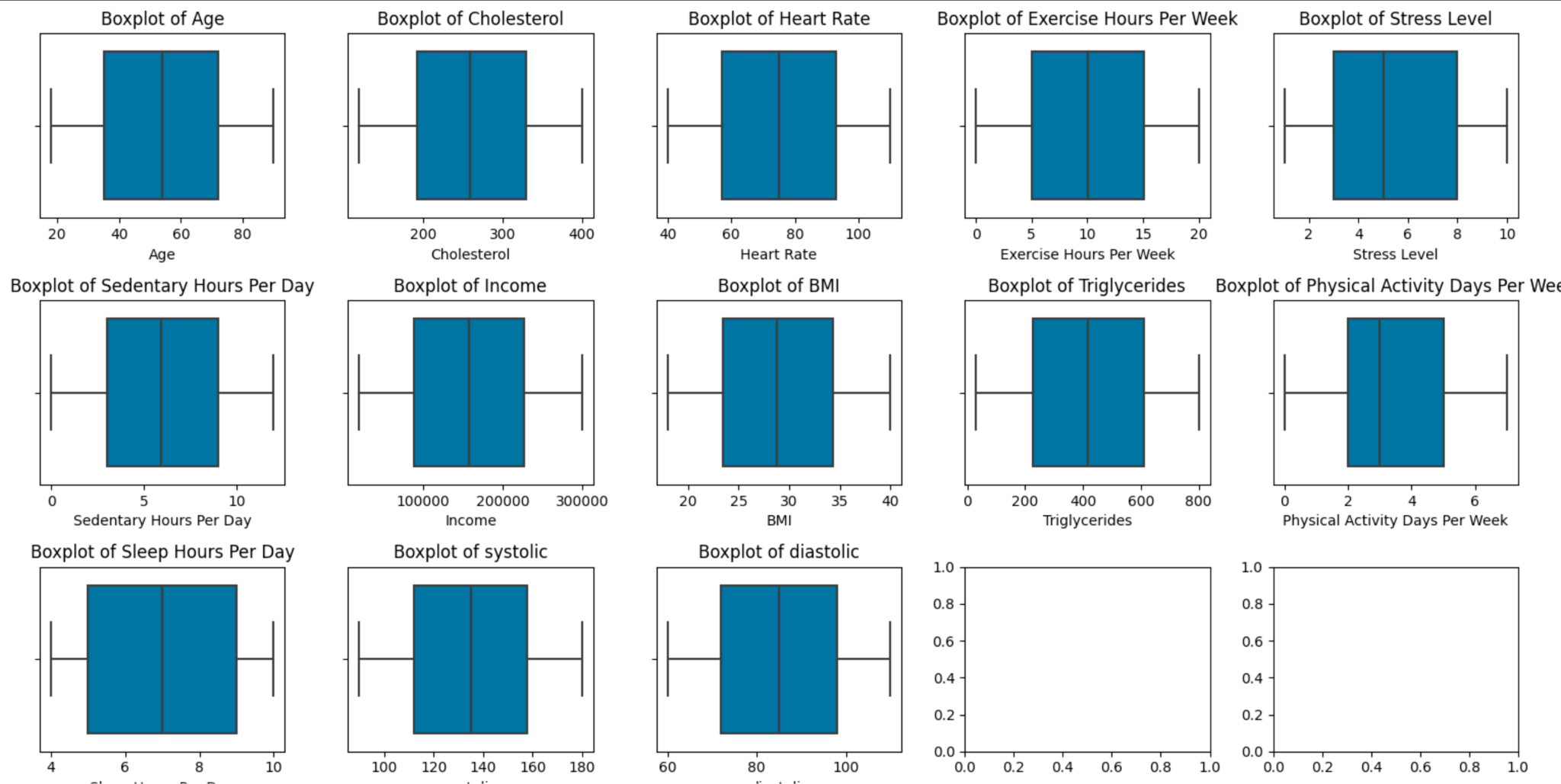


Fig.6 Outlier analysis - Boxplot

**Exploration Data Analysis**

**Label Encoding**

for the numeric features have medical explanations, such as “Blood Pressure”, “Triglycerides”, “Cholesterol”, “BMI” and “Heart Rate”. We do data expansion by referring to CDC’s (Centers for Disease Control and Prevention: <https://cdc.gov>) explanations and mapping numeric data to human readable levels according to the guidelines CDC provides. We also labeled features like “Sex” and “Hemisphere”, which have Object datatype to boolean type in new columns.

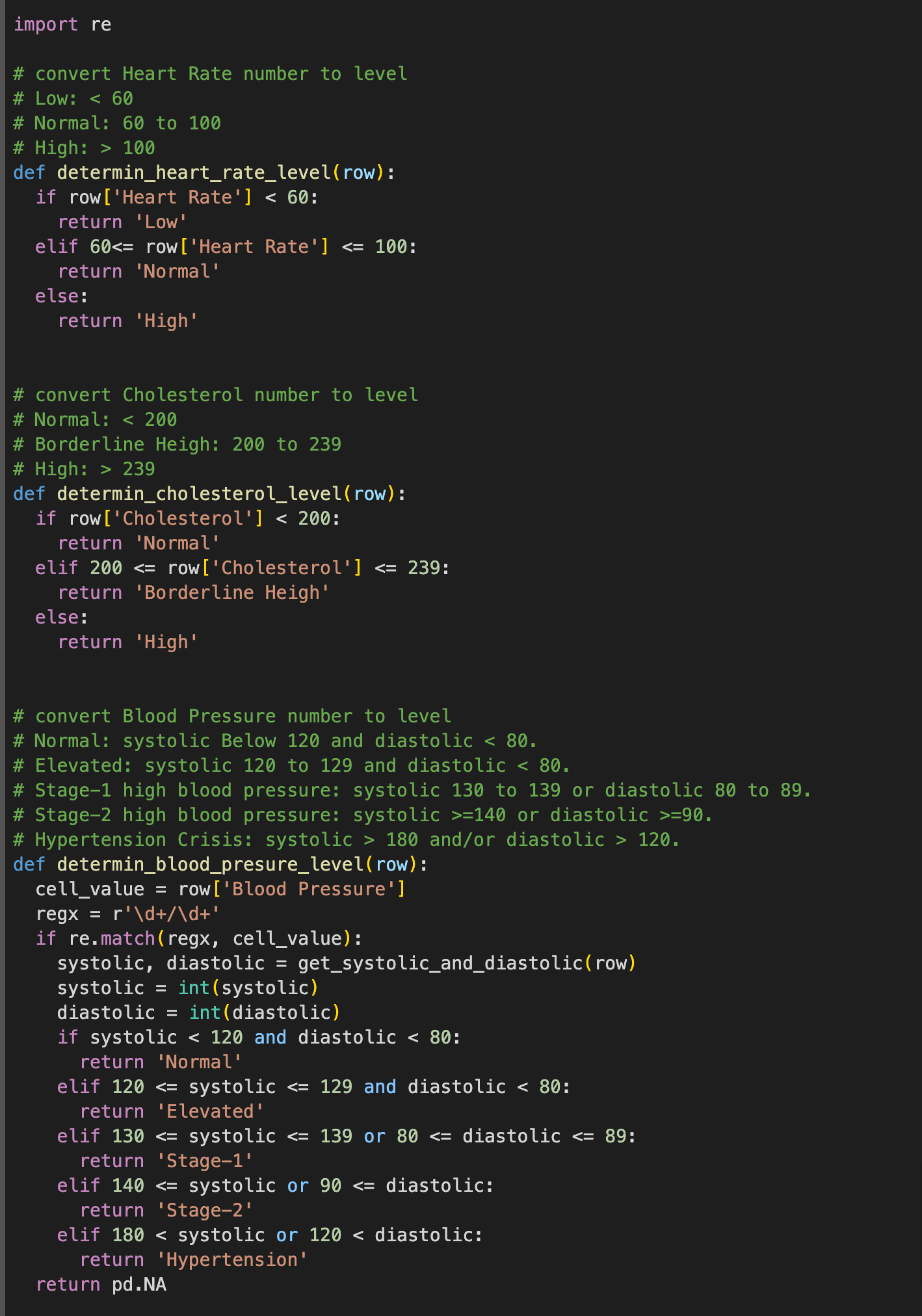


Fig.7 Lable Encoding for Heart Rate, Cholesterol and Blood Pressure

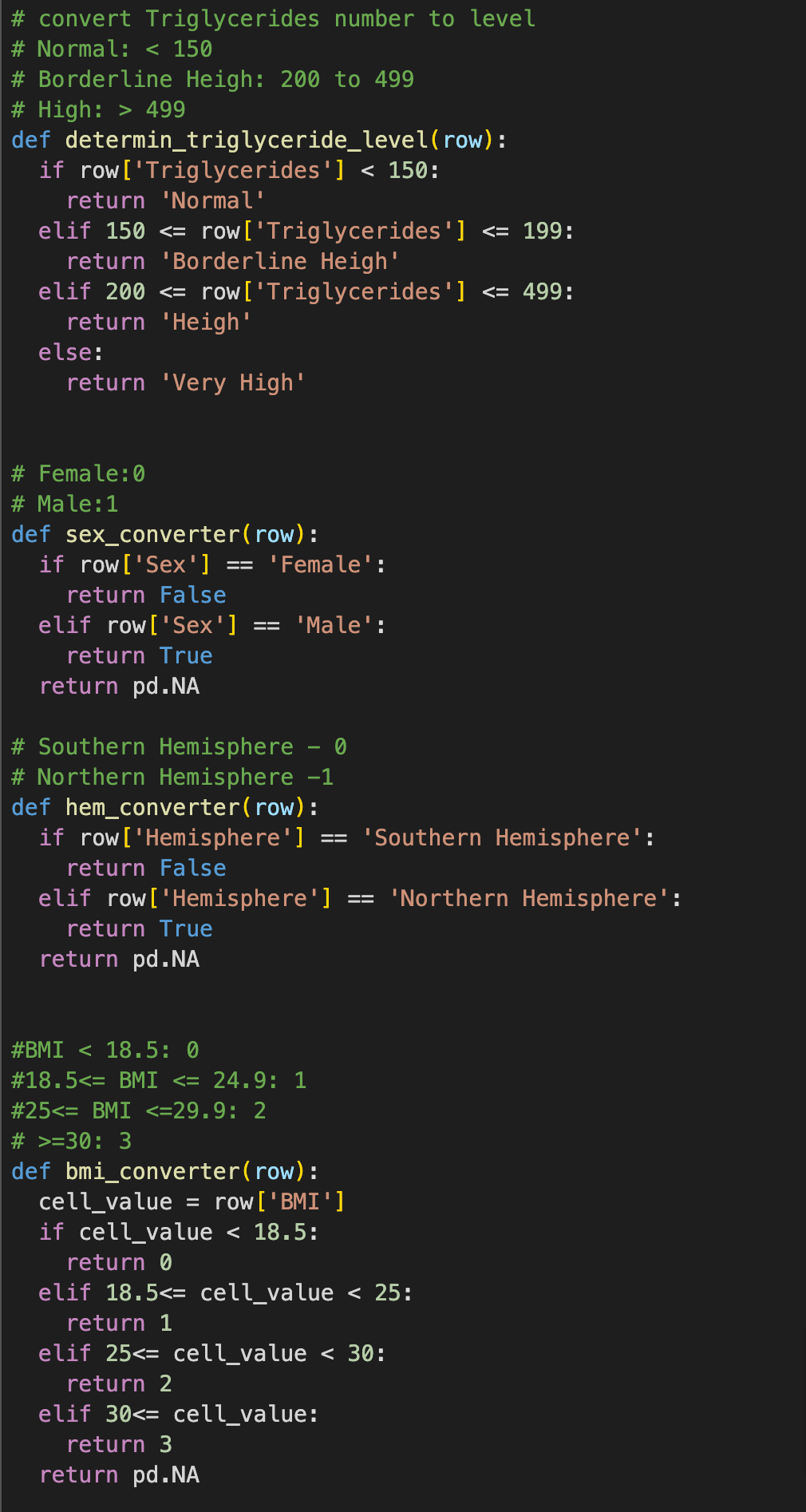


Fig.8 Lable Encoding for Triglyceride, Sex, Hemisphere and BMI

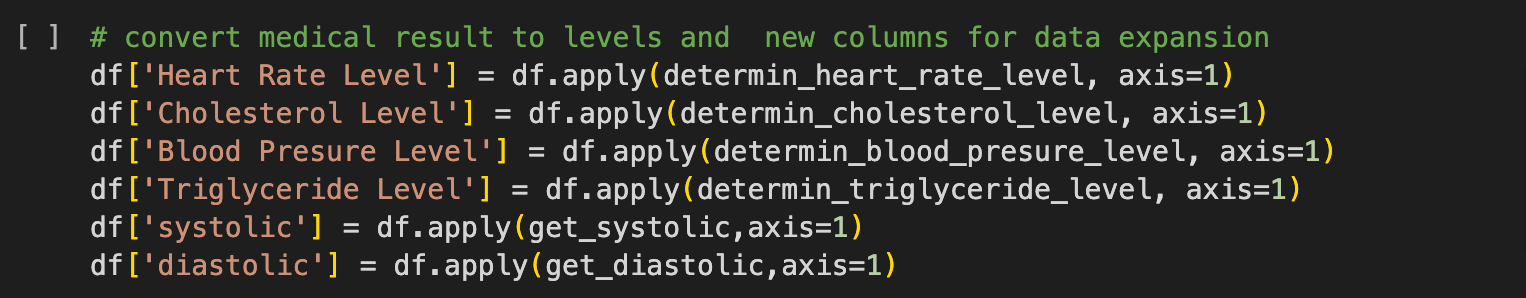


Fig.9 Data Expansion

Drop the translated column and PatientID, which is useless.

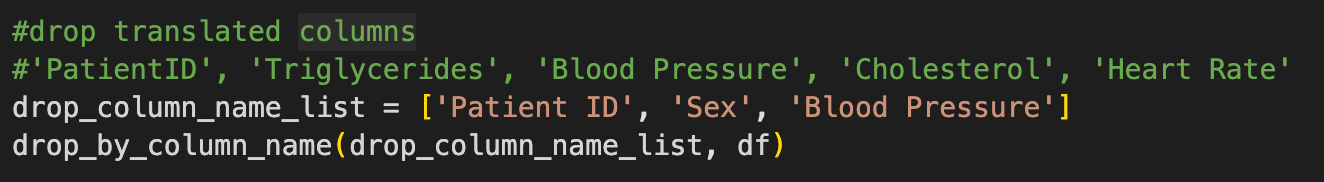


Fig.10 Drop column

**Correlation Matrix**

We use the Seaborn library to discover structures in heatmap data after data cleaning and labeling.

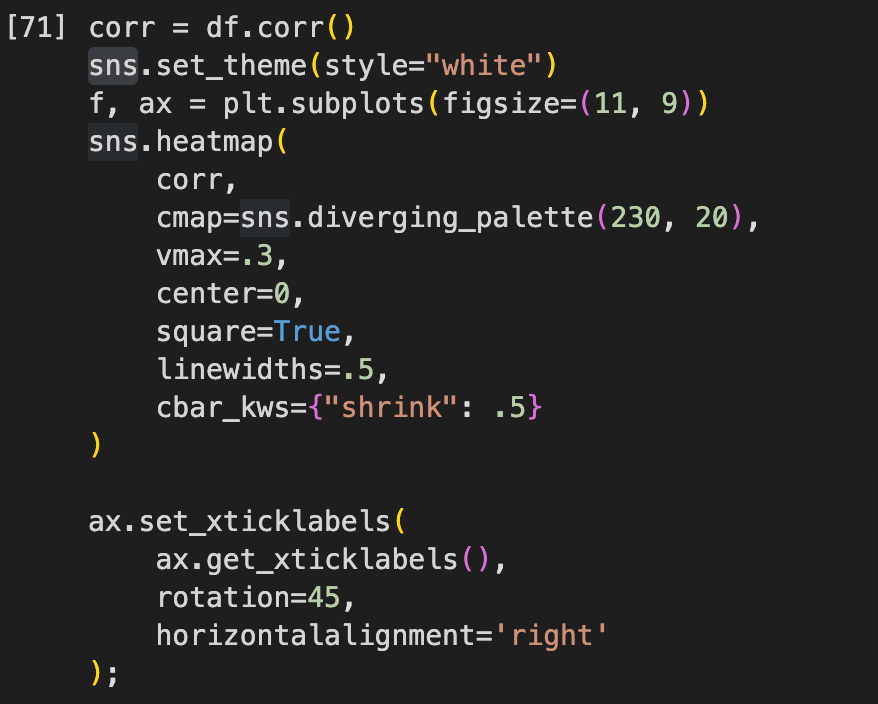


Fig.11 Code of correlation matrix



Fig.12 Correlation Matrix

But the limited conclusion we can get according to the matrix. So we continued exploring more data with graphs, and we found Heart Attack Risk has a high correlation with Cholesterol Level, Blood Pressure Level, and Triglyceride Level.

Sample codes to build the heatmap with Heart Attack Risk by Country and Cholesterol Level are as follows. The other two are almost the same except for the variables.

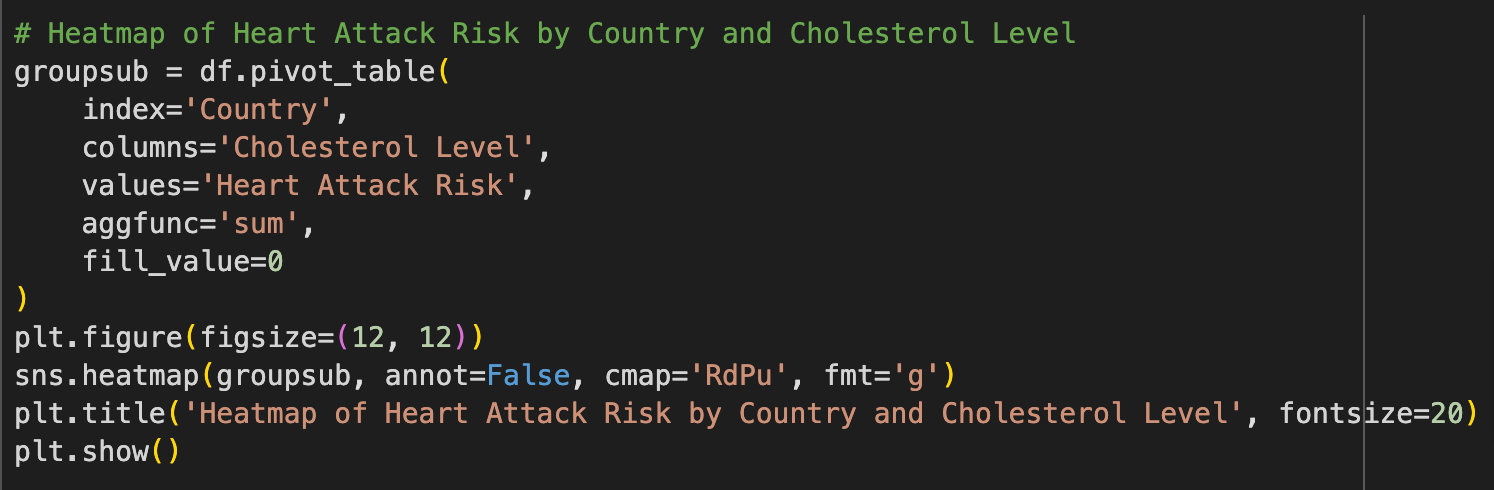
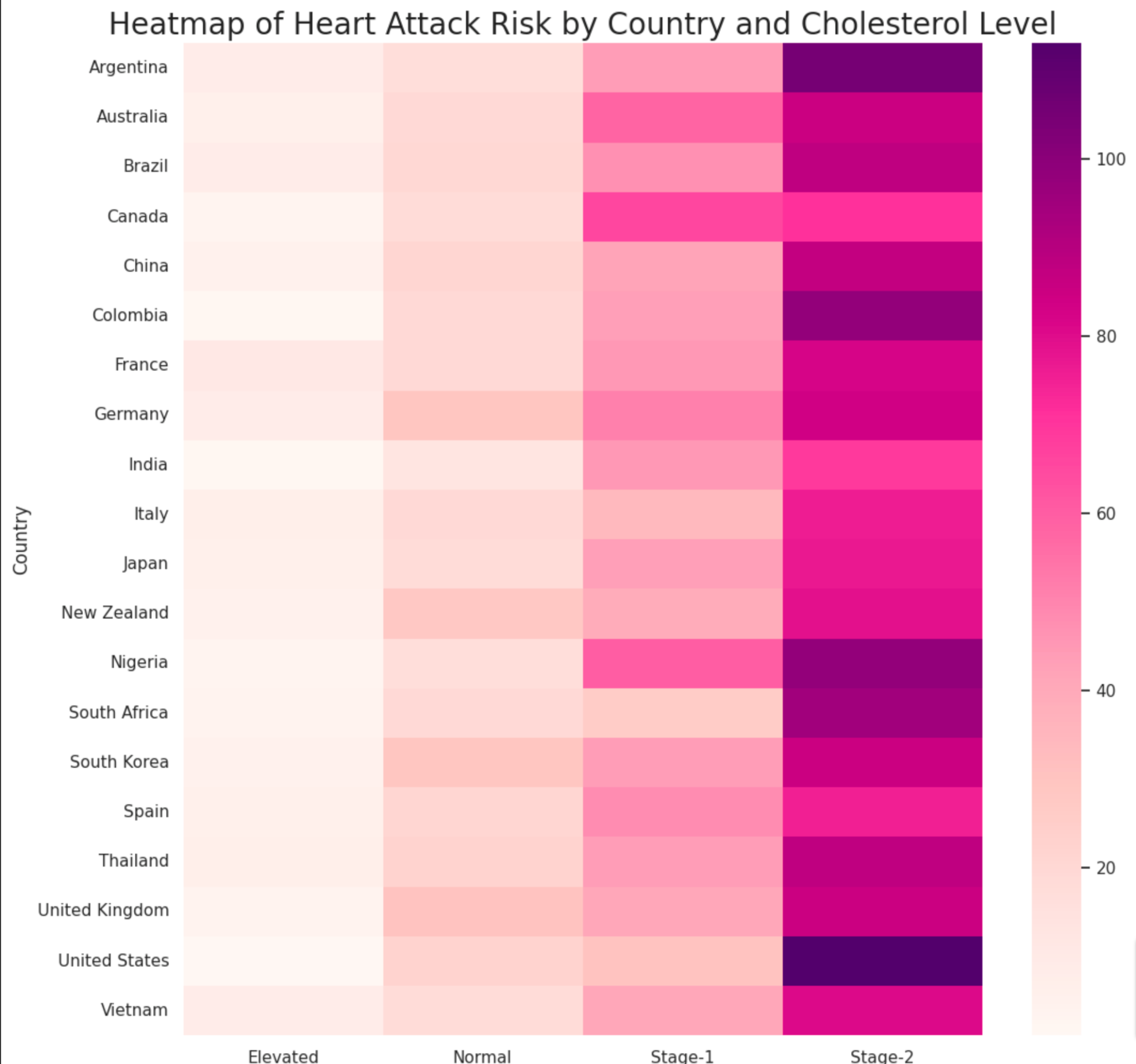


Fig.13 Code of Headmap



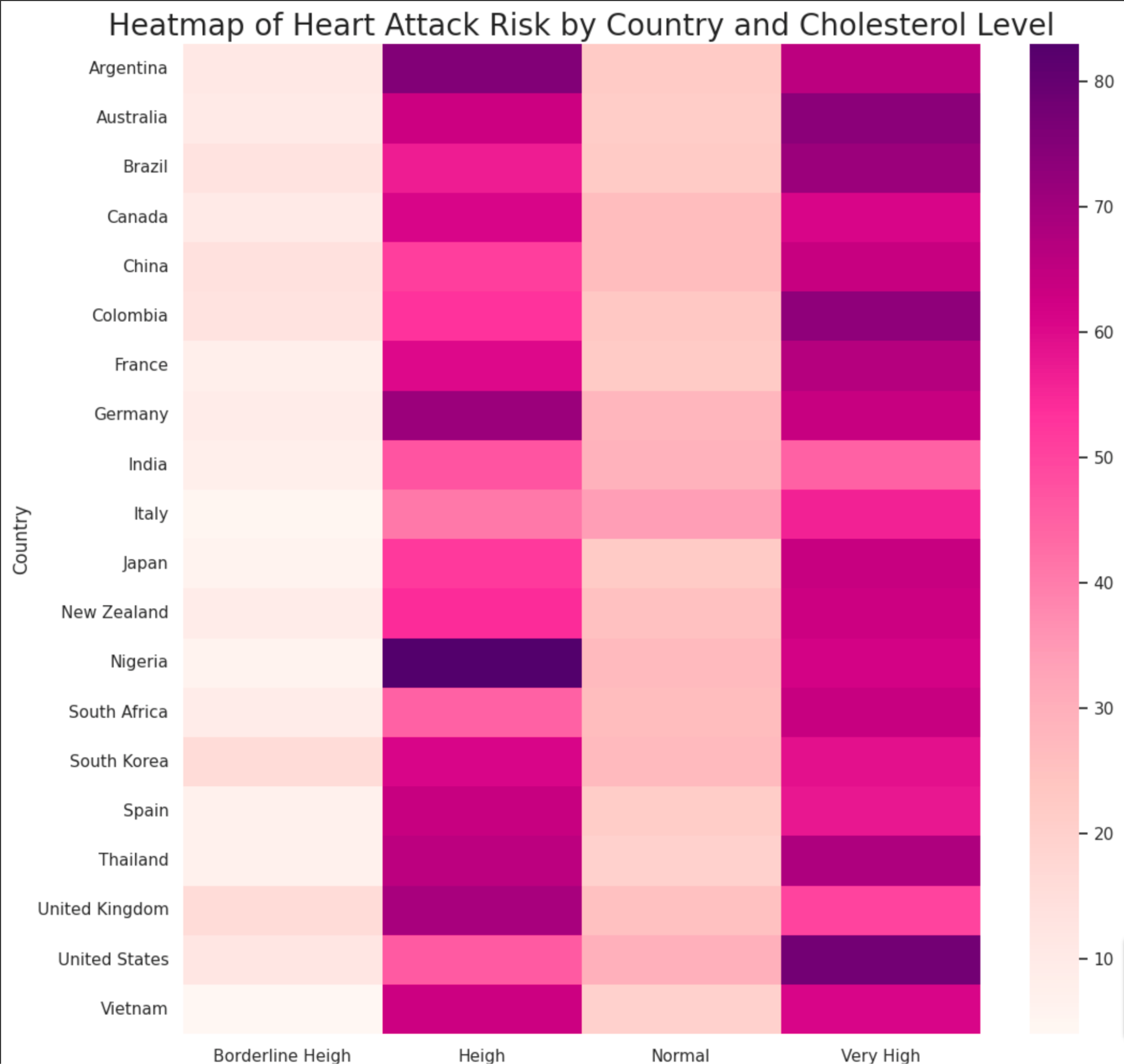


Fig.14 The Heat Maps and the correlation insights we can get from

We explored the variable “BMI” with “Diet” together with Cholesterol Level, Blood Pressure Level, and Triglyceride Level to explore correlations with “Heart Attack Rist” by using “seaborn.relplot”. However, we can not get any correlation between them. Take one exploration for example

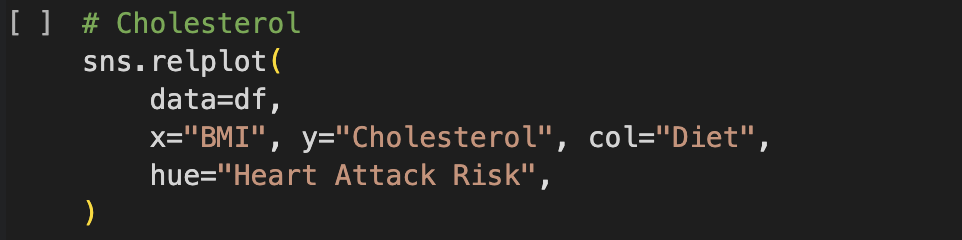


Fig.15 Sample code of relplot

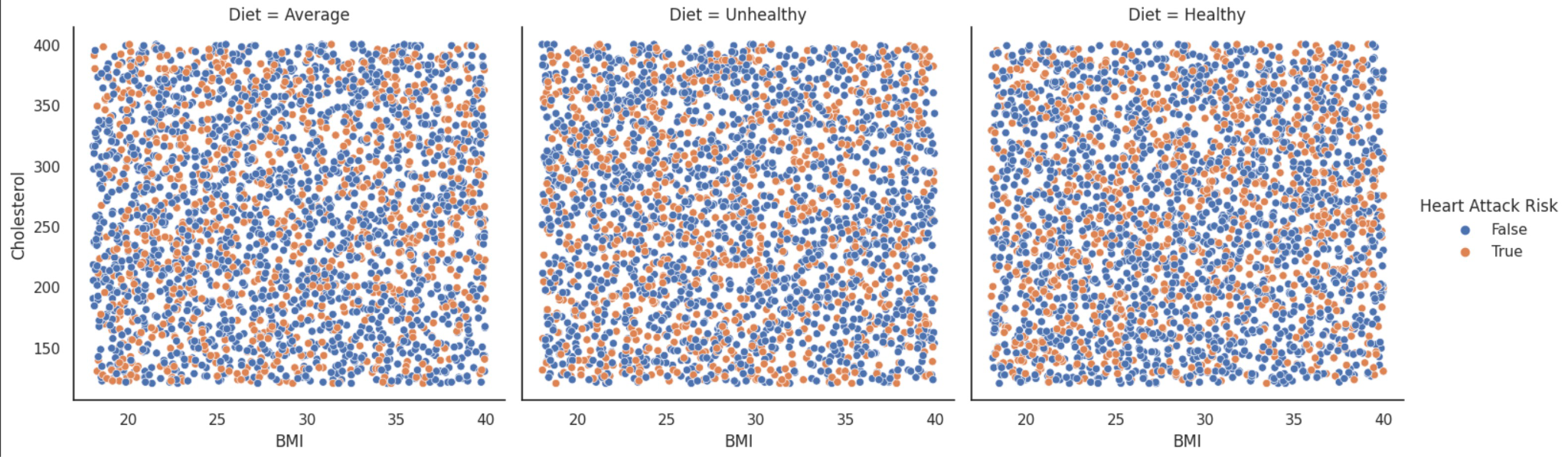


Fig.16 Sample relplot graph

We also explore geographic data analysis to see the distribution of heart attack risk at the Country level and country level, by using “plotly.express”. The choropleth map illustrates regional disparities in heart attack risks, with countries like Argentina and Nigeria exhibiting higher counts, pointing to significant public health challenges in these nations, whereas New Zealand and Vietnam show lower counts, suggesting fewer incidences or better management of heart attack risks. The data indicates a diverse spectrum of heart attack risk counts in Asia, with China showing a higher risk and Vietnam a lower one, while European countries display moderate levels, which could reflect their healthcare infrastructure and lifestyle factors.



Fig.17 Code of plotly.express.choropleth

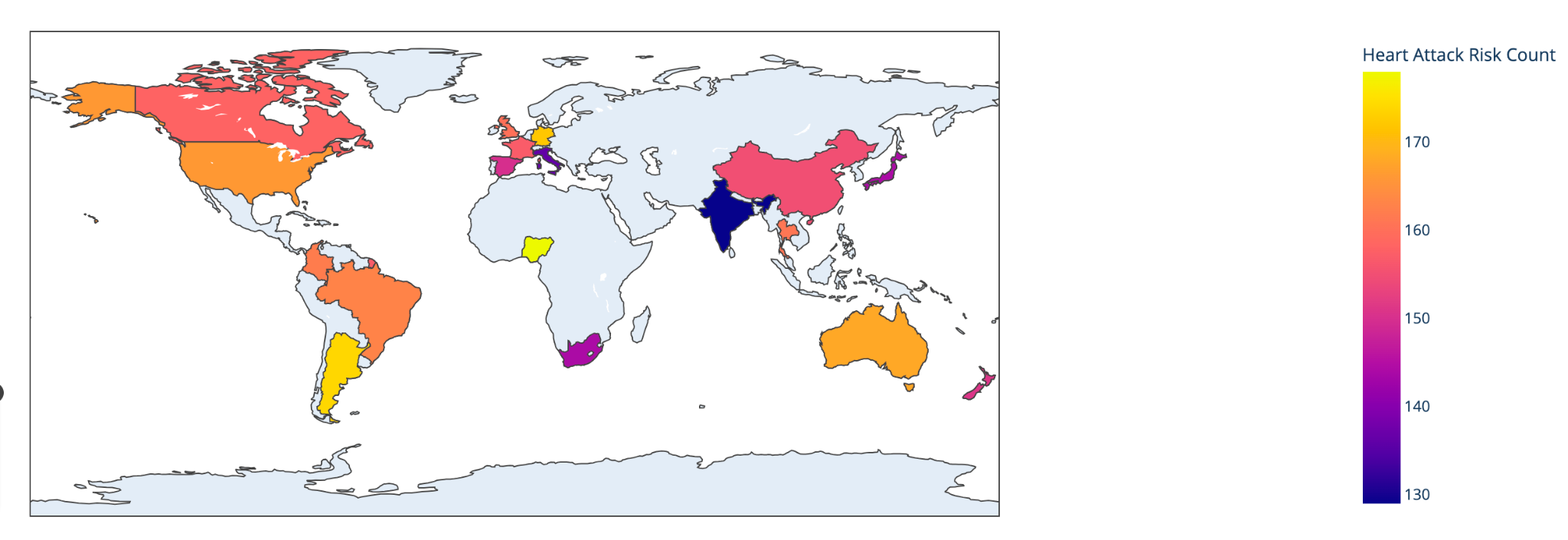


Fig.18 Heatmap show positive heart attack risk worldwide

The next choropleth map Fig 20 presents a global overview of high cholesterol counts (>= 240), with countries color-coded according to their respective levels. Notable observations include France with the highest count at 274, suggesting a significant public health concern, and India with a lower count of 227, which may reflect differences in diet, lifestyle, or health policy effectiveness.

There is a notable variation in high cholesterol counts among the depicted countries, with European nations like France and Germany showing some of the highest levels, while Asian countries display a range with South Korea on the lower end at 234 and China with a moderately high count of 243, indicating diverse health profiles across these regions.



Fig.19 Code of pie chart

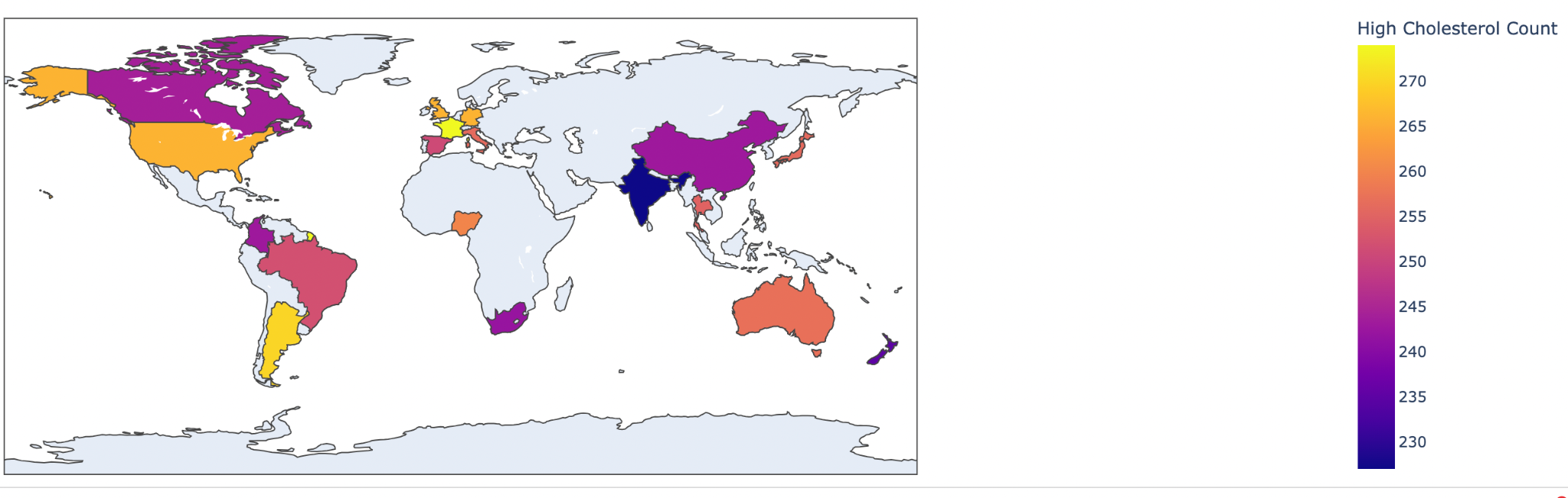


Fig.20 High cholesterol count (>=240) worldwide

A pie chart that shows the heart attack risk in each continent, where Asia has the highest relative heart attack risk percentage at 28.7%, which could be attributed to its large population and regional lifestyle and health factors. Europe also has a considerable proportion of heart attack risk at 24.7%, potentially related to lifestyle choices and an aging population, while other continents like Africa, North America, Australia, and South America show more evenly distributed risk percentages.

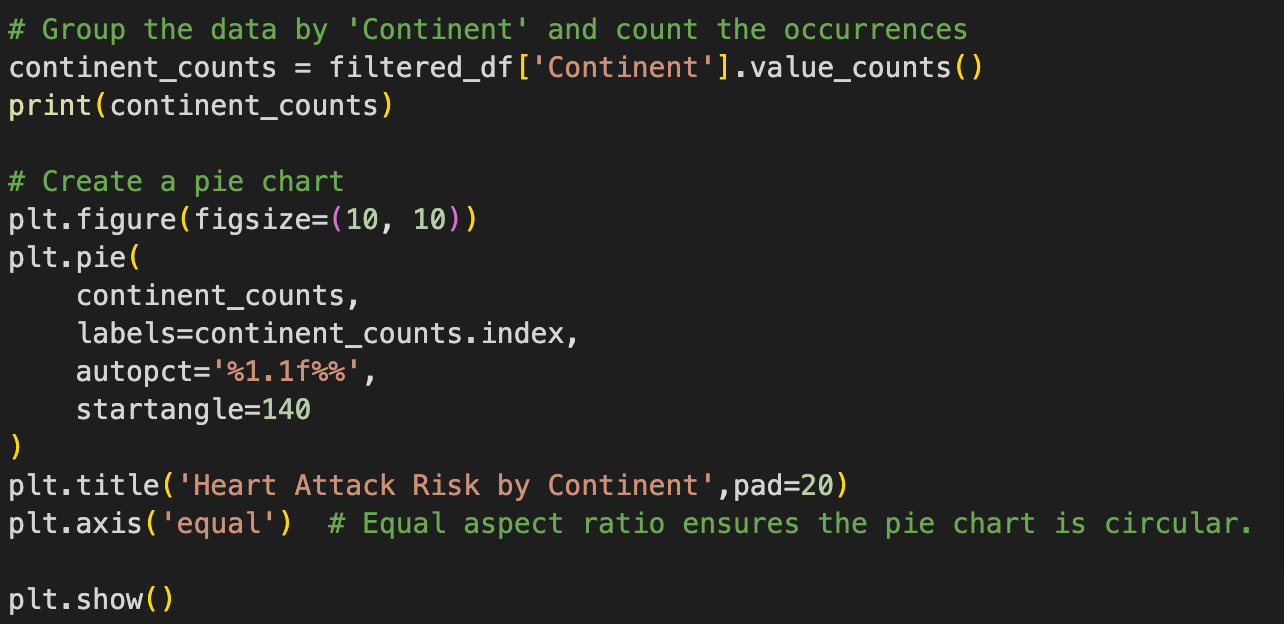


Fig. 21 Code of pie chart

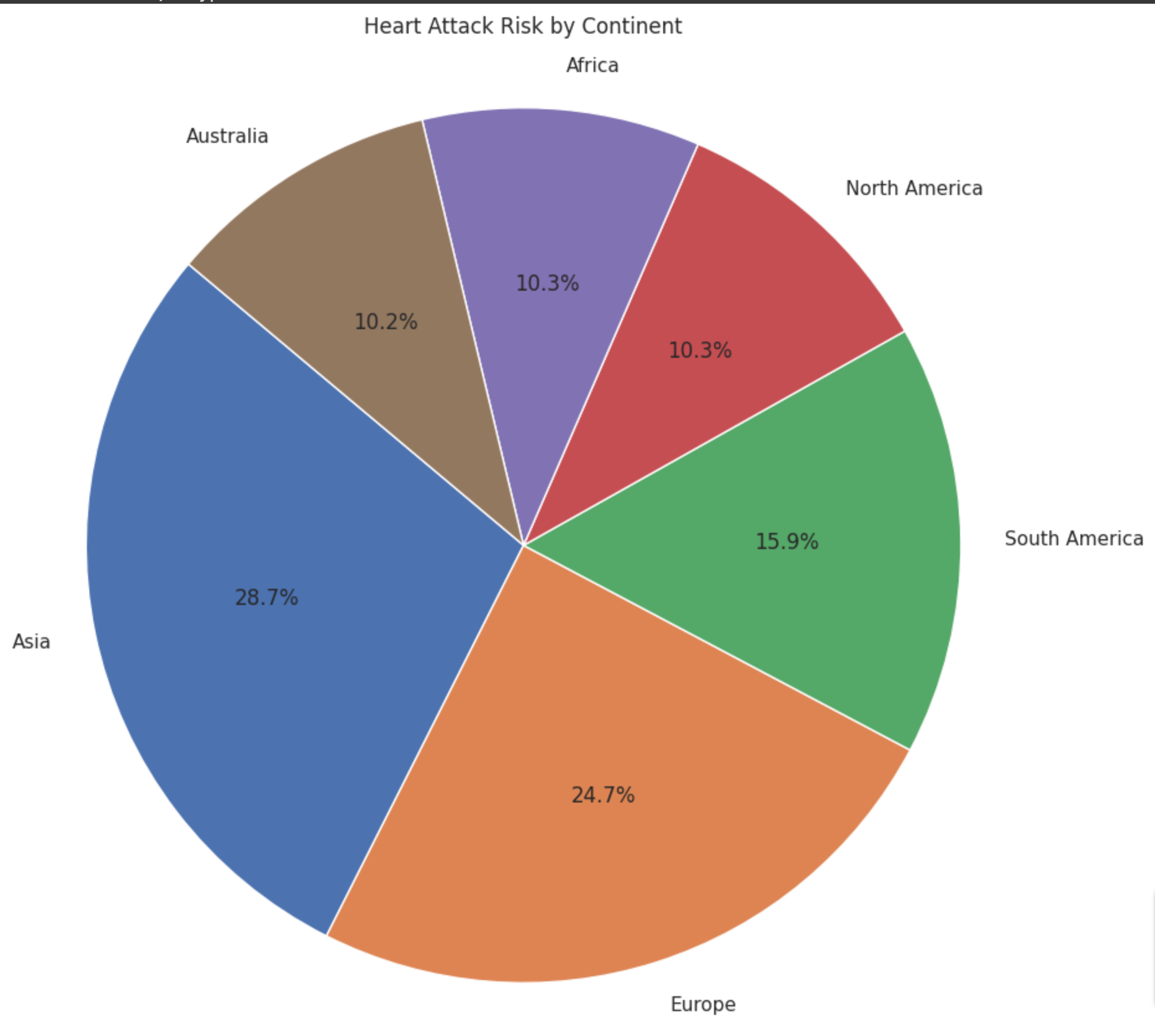


Fig. 22 Pie chart analysis by Continent

**What is next?**

**Data exploration and correlation analysis**

We have already found some correlations between Heart Attack Rate and Cholesterol Level, Blood Pressure Level, and Triglyceride Level. Next, we will continue with data visualization, to create histograms, scatter plots, and any other plots to get a sense of the data distribution and relationships.

Exploring summary statistics is also a good way to get insights into different features to understand their central tendencies.

**Feature Engineering**

At this point, we have several ideas, beyond using a single attribute, we plan to combine two or several together. For example, a cholesterol-to-age ratio, because older people tend to have different cholesterol thresholds; normalizing cholesterol by age could potentially provide a more nuanced feature.

**Machine learning models**

Split the dataset into training and testing sets and select the appropriate machine-learning models. Such as Decision Trees, Random Forest, SVM, etc. Evaluate their performance with currency, precision, recall, and F1-score.

**Hyperparameter tuning**

Identify the hyperparameters associated with our models that need tuning with range for each hyperparameter.

**Cross-validation**

To evaluate model performance during hyperparameter tunning. And model performance comparison.

After that, we will summarize our findings and provide insights into which factors are most predictive of heart attack risk and any other relevant discoveries.

**References**

*Heart attack Risk Prediction Dataset*. (2023, September 27). <https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset>

Fletcher, J. (2023, June 30). *What should my cholesterol level be at my age?* <https://www.medicalnewstoday.com/articles/315900#summary>

Aayush Malik. (2020). *9 Steps for Solving Data Science Problems.* Towards Data Science

*Understanding blood pressure readings*. (2023, October 17). www.heart.org. <https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings>

*About high cholesterol | cdc.gov*. (2023, March 20). Centers for Disease Control and Prevention. <https://www.cdc.gov/cholesterol/about.htm>