Mid-Term project Report

Heart Attack Data Analysis

Harsha Joshi
hj2468 / N17073170
MSCS

Disclaimer: This report is submitted to NYU's Computer Science Department for the sole purpose of assigning an internship grade. The information remains confidential and proprietary to the company.

NYU

Contents

Introduction and Background

Heart attacks, also known as myocardial infarctions, are a leading cause of death worldwide. They are a critical issue that has become more common now than ever due to the many lifestyle changes that humans have adopted over the years. These lifestyle changes may include a lack of exercise, smoking, alcohol consumption, dieting habits, and many more.

It is important that proper analysis be done on different reasons for heart attacks so that people can be made aware of some precautions they can start taking for their health.

Some other reasons why data analysis on heart attacks should be done are as follows:

1. They result in a significant number of fatalities and contribute to a substantial burden of disease, impacting both individuals and society.
2. Heart attacks typically come on suddenly and can be quite severe, demanding immediate medical attention.
3. The sudden and intense nature of the symptoms can be life-threatening.
4. Around the world, governments and healthcare organizations make significant investments in public health campaigns.
5. These efforts aim to increase awareness about the risk factors, symptoms, and preventive measures associated with heart attacks.

Improving patient care will help us understand the trend and detect early causes of it.

Using predictive models can assist in anticipating how many patients might need care, when disease outbreaks could occur, and what resources hospitals might require. This way, healthcare facilities can get ready ahead of time to handle any sudden increases in demand effectively.

I got the inspiration to pick this topic for mid-term project after reading a blog post called "COVID-19 Surges Linked to Spikes in Heart Attacks". This blog talks about how the heart attack rate has gone up after Covid-19 and what the supposed reasons are for this change.

Although I have not used the covid-19 data set that they have used, I made an effort to understand what the heart attack patterns are in a few countries around the world and what the reasons behind them are.

Dataset

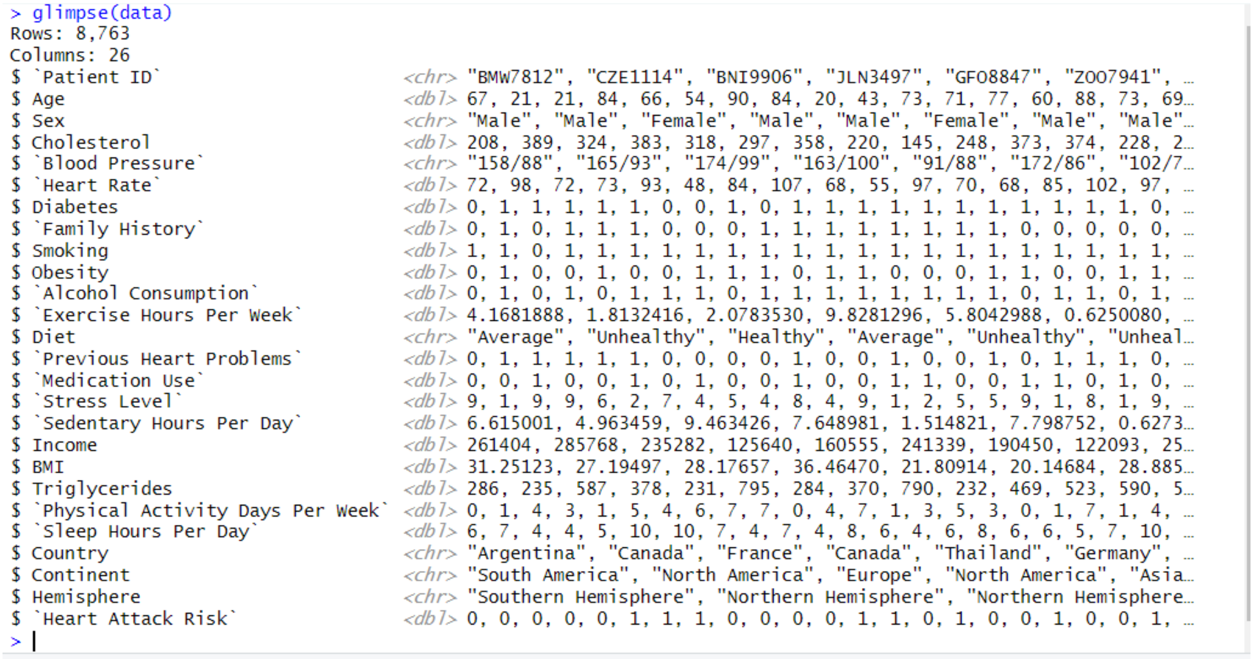
I used Kaggle to grab a dataset to work on; the following is the link to it:
https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset
Author: Sourav Banerjee

The dataset name is Health Attack Prediction Dataset.
The data seems to have been created for the purpose of doing a predictive analysis on the reasons in different countries that are causing heart attacks

The dataset required no cleaning; it didn't require any data wrangling option.

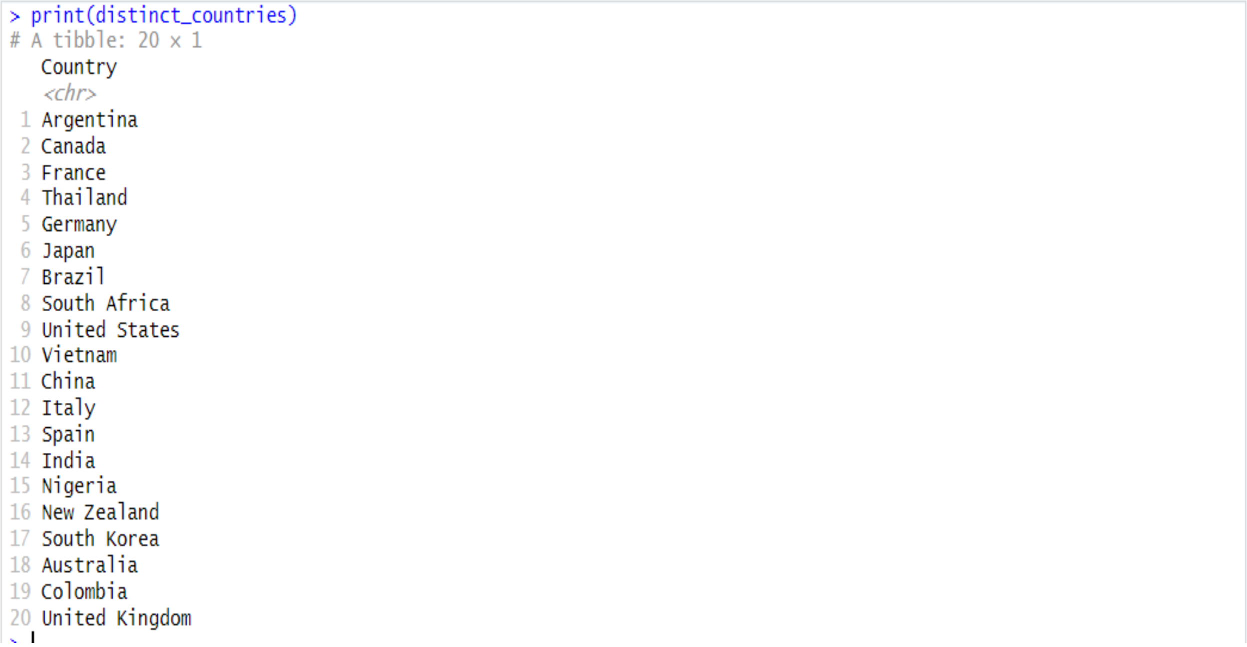
The data set has around 8764 entries and 26 columns

Following are the different columns and their data types:



```
> glimpse(data)
Rows: 8,763
Columns: 26
$ `Patient ID`                      <chr> "BMW7812", "CZE1114", "BNI9906", "JLN3497", "GFO8847", "ZOO7941", …
$ Age                               <dbl> 67, 21, 21, 84, 66, 54, 90, 84, 20, 43, 73, 71, 77, 60, 88, 73, 69…
$ Sex                               <chr> "Male", "Male", "Female", "Male", "Male", "Female", "Male", "Male"…
$ Cholesterol                       <dbl> 208, 389, 324, 383, 318, 297, 358, 220, 145, 248, 373, 374, 228, 2…
$ `Blood Pressure`                  <chr> "158/88", "165/93", "174/99", "163/100", "91/88", "172/86", "102/7…
$ `Heart Rate`                      <dbl> 72, 98, 72, 73, 93, 48, 84, 107, 68, 55, 97, 70, 68, 85, 102, 97, …
$ Diabetes                          <dbl> 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, …
$ `Family History`                  <dbl> 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, …
$ Smoking                           <dbl> 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ Obesity                           <dbl> 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1, …
$ `Alcohol Consumption`             <dbl> 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, …
$ `Exercise Hours Per Week`         <dbl> 4.1681888, 1.8132416, 2.0783530, 9.8281296, 5.8042988, 0.6250080, …
$ Diet                              <chr> "Average", "Unhealthy", "Healthy", "Average", "Unhealthy", "Unheal…
$ `Previous Heart Problems`         <dbl> 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, …
$ `Medication Use`                  <dbl> 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, …
$ `Stress Level`                    <dbl> 9, 1, 9, 9, 6, 2, 7, 4, 5, 4, 8, 4, 9, 1, 2, 5, 5, 9, 1, 8, 1, 9, …
$ `Sedentary Hours Per Day`         <dbl> 6.615001, 4.963459, 9.463426, 7.648981, 1.514821, 7.798752, 0.6273…
$ Income                            <dbl> 261404, 285768, 235282, 125640, 160555, 241339, 190450, 122093, 25…
$ BMI                               <dbl> 31.25123, 27.19497, 28.17657, 36.46470, 21.80914, 20.14684, 28.885…
$ Triglycerides                     <dbl> 286, 235, 587, 378, 231, 795, 284, 370, 790, 232, 469, 523, 590, 5…
$ `Physical Activity Days Per Week` <dbl> 0, 1, 4, 3, 1, 5, 4, 6, 7, 7, 0, 4, 7, 1, 3, 5, 3, 0, 1, 7, 1, 4, …
$ `Sleep Hours Per Day`             <dbl> 6, 7, 4, 4, 5, 10, 10, 7, 4, 7, 4, 8, 6, 4, 6, 8, 6, 6, 5, 7, 10, …
$ Country                           <chr> "Argentina", "Canada", "France", "Canada", "Thailand", "Germany", …
$ Continent                         <chr> "South America", "North America", "Europe", "North America", "Asia…
$ Hemisphere                        <chr> "Southern Hemisphere", "Northern Hemisphere", "Northern Hemisphere…
$ `Heart Attack Risk`               <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, …
> |
```

The dataset is spread over various countries. Those are as follows:

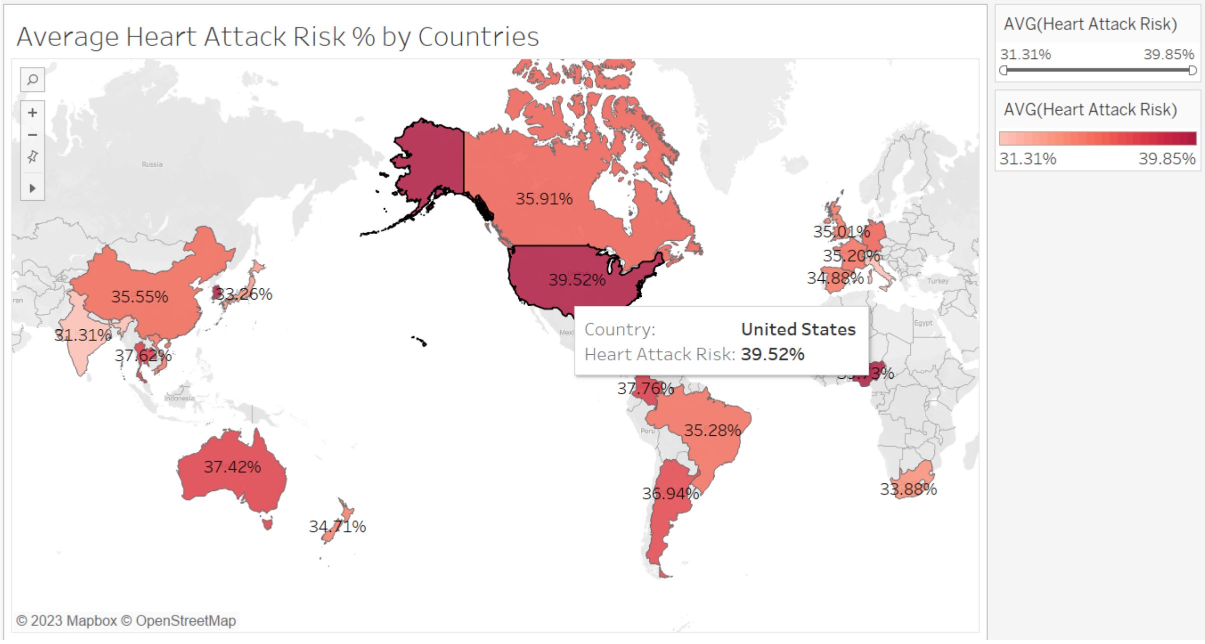


```
> print(distinct_countries)
# A tibble: 20 x 1
   Country
   <chr>
 1 Argentina
 2 Canada
 3 France
 4 Thailand
 5 Germany
 6 Japan
 7 Brazil
 8 South Africa
 9 United States
10 Vietnam
11 China
12 Italy
13 Spain
14 India
15 Nigeria
16 New Zealand
17 South Korea
18 Australia
19 Colombia
20 United Kingdom
>
```

3. Data Story

Chart 1: Map

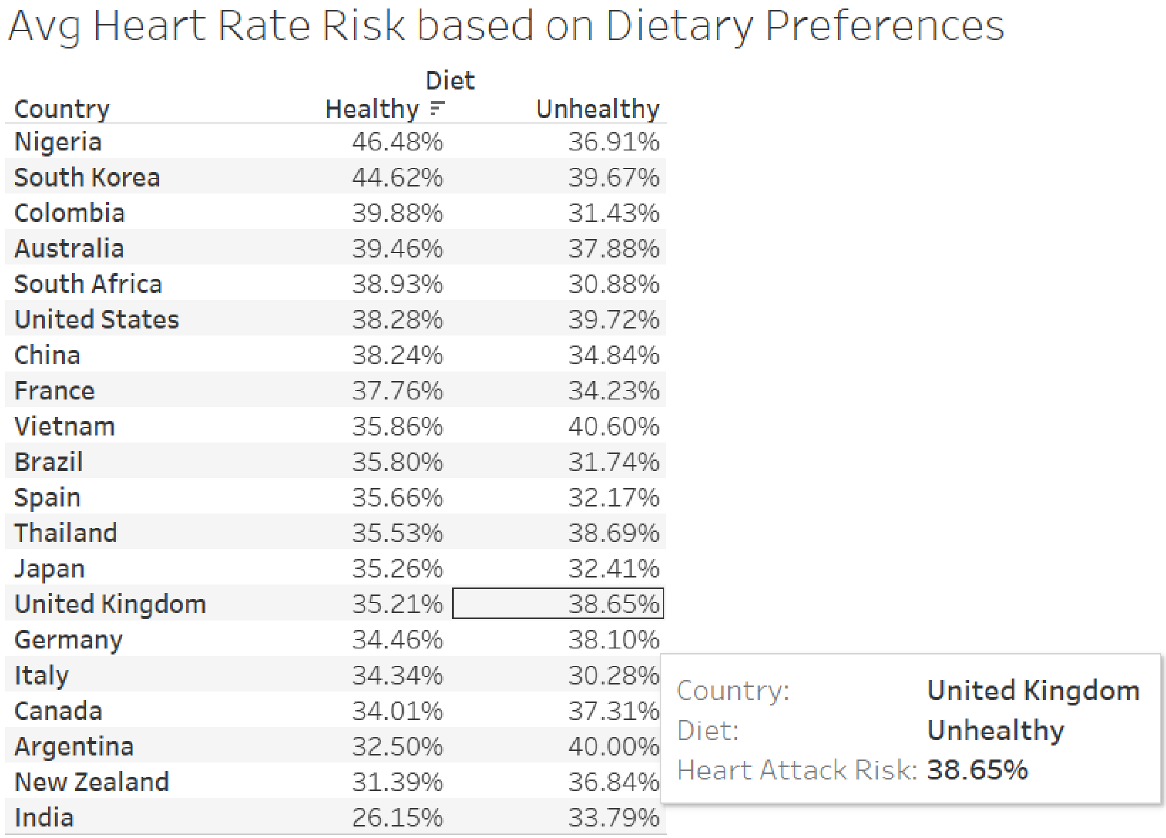
I started by analyzing the highest average heart attack percentage in each country and understanding the trend.
I chose the country variable because different countries have different lifestyle patterns and observing them can give a good insight into the reasons for heart attacks.



Country: United States
Heart Attack Risk: 39.52%

Using a map, we can follow the color ombre trend to understand the average heart attack rates, going from highest to lowest, from darkest to lightest shade of red.
Based on the map, we can conclude that heart attack rates in South Korea, Nigeria, and the United States are the countries with the highest risk, and Italy, Japan, and India have the lowest.
I used the red color because that color signifies danger/alert and I thought it would be appropriate to show the heart attack risks. Using the rates, I created an ombre effect of red so that, just looking at the visualization, the user can understand which countries are on the higher end and which are on the lower end. I have also put rates on the label, giving more transparency to the reader
In further charts, let's understand what may be the cause of the above results and what factors can determine what the people of a country should do to decrease their chances of heart attack.

Chart 2: Table



Avg Heart Rate Risk based on Dietary Preferences

	Diet	
Country	Healthy	Unhealthy
Nigeria	46.48%	36.91%
South Korea	44.62%	39.67%
Colombia	39.88%	31.43%
Australia	39.46%	37.88%
South Africa	38.93%	30.88%
United States	38.28%	39.72%
China	38.24%	34.84%
France	37.76%	34.23%
Vietnam	35.86%	40.60%
Brazil	35.80%	31.74%
Spain	35.66%	32.17%
Thailand	35.53%	38.69%
Japan	35.26%	32.41%
United Kingdom	35.21%	38.65%
Germany	34.46%	38.10%
Italy	34.34%	30.28%
Canada	34.01%	37.31%
Argentina	32.50%	40.00%
New Zealand	31.39%	36.84%
India	26.15%	33.79%

Country: United Kingdom
Diet: Unhealthy
Heart Attack Risk: 38.65%

The above table gives us insight into the rate of heart attacks in different countries, given their diet preferences. The data set has three diets: healthy, unhealthy, and average. The above charts only showcase healthy and unhealthy diets.

The table shows us in every country what the rate of heart attack is given a healthy and unhealthy diet.

While I understand that basing conclusions on only one parameter in this case, diet, is not ideal, it can suggest a few things.

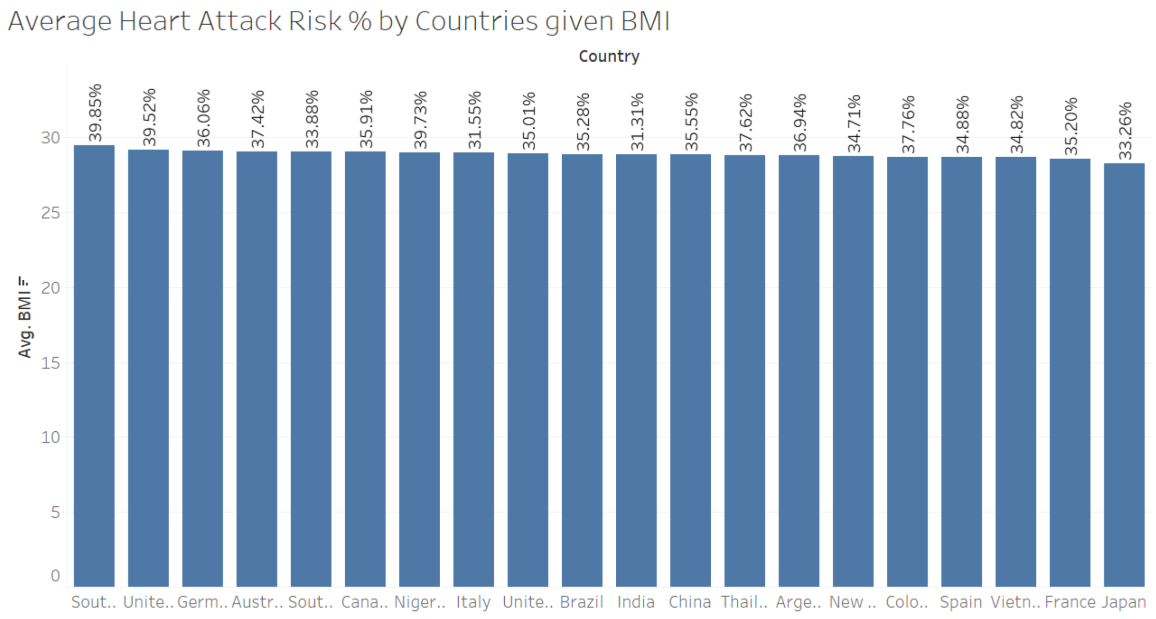
From the above chart, we can conclude:

1. In countries like Argentina and India, healthy eating population has lesser chance of heart attack by almost 8-10%.
2. In countries like the USA and Australia, the difference in the likelihood of getting a heart attack differs by around 1.5%, which is almost negligible. There might be another reason or cause of heart attacks in this country but diet is not one of them.

The best practices used in this table are to give the heading of the chart and also rename the sheet name.

The average heart attack risk for a healthy diet is arranged in descending order for the reader to be able to understand the table better and comprehend the metrics.

Chart 3: Bar Chart

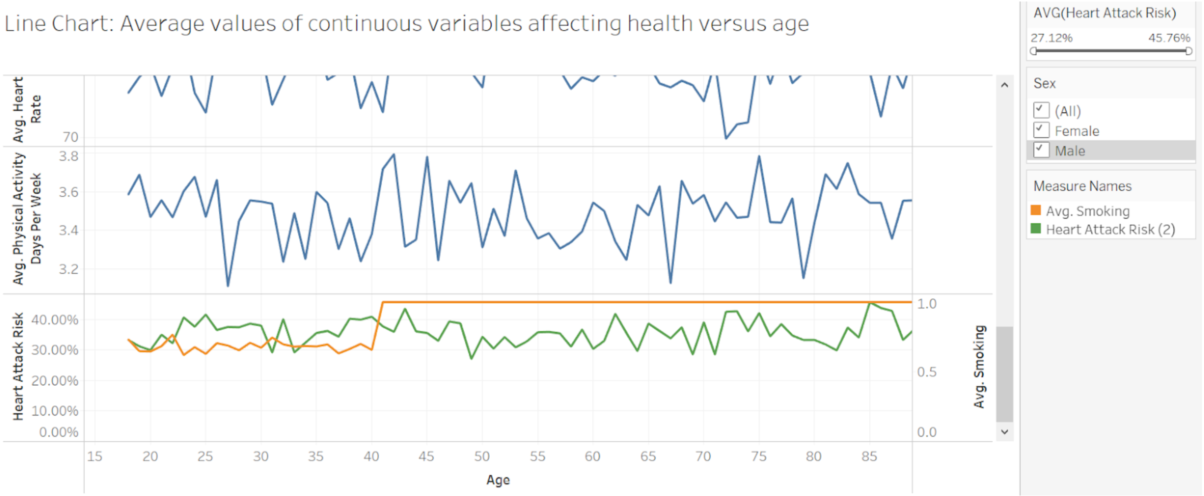


We understand the above chart is a representation of countries with the highest to lowest BMI, with a label of that country's associated average heart attack risk rates.

From Chart 2, we saw that there are other parameters than diet that are responsible for heart attacks in countries like the USA and Australia. In this chart, we can see they have a high BMI, which can be a reason for their high heart attack rates.

The best practices that I have used are arranging the chart from highest to lowest, naming the axes properly, and maintaining a uniform color.

Chart 4: Line Chart



AVG(Heart Attack Risk)
27.12% 45.76%

Sex
☑ (All)
☑ Female
☑ Male

The above line chart shows that something is off with the dataset and it is not quite close to our expectations. One would expect that as age increases, the average fitness of that age group will decrease, or that as age increases, the average BMI might increase. Similarly, for other variables as well. However, in our dataset, that doesn't happen.

I have done the analysis for several variables:

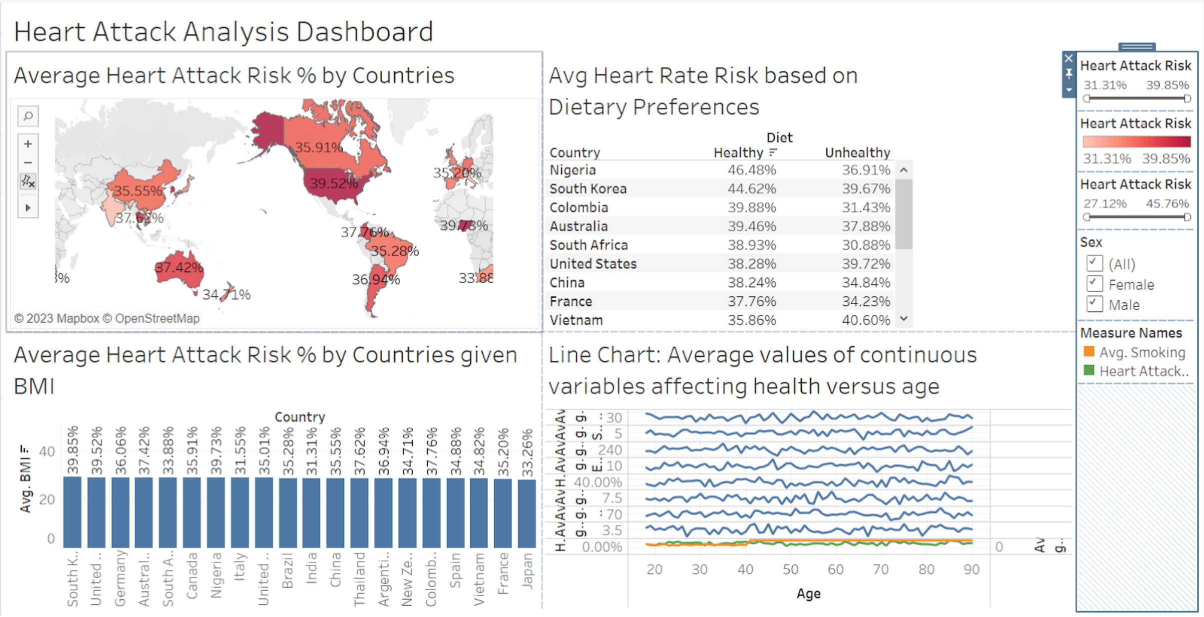
1. Exercise hours per week
2. BMI
3. Sedentary Hours per day
4. Cholesterol Levels
5. Heart Attack Risk
6. Smoking
7. Sleeping Hours
8. Stress Levels, etc.

I observed that there is no specific trend in averages of all of these variables versus age; that is, neither they increase nor decrease in an overall pattern.

Since our ultimate goal is to analyze heart rate risk and there is a very high correlation between heart attack risk and smoking, I made a dual axis graph to identify any trends. However, I noticed that in our dataset, every individual who is above the age of 41 is a smoker, which is why in the chart above, the average smoking line became constant. Also, there is no specific pattern between heart attack risk and increasing age.

The best practices that I have used while creating this chart are to name the chart and sheet appropriately, named axes to give the reader a proper context and in the line chart with dual axis, separately colored lines.

Dashboard:



Heart Attack Analysis Dashboard

Average Heart Attack Risk % by Countries

Avg Heart Rate Risk based on Dietary Preferences

	Diet	
Country	Healthy	Unhealthy
Nigeria	46.48%	36.91%
South Korea	44.62%	39.67%
Colombia	39.88%	31.43%
Australia	39.46%	37.88%
South Africa	38.93%	30.88%
United States	38.28%	39.72%
China	38.24%	34.84%
France	37.76%	34.23%
Vietnam	35.86%	40.60%

Heart Attack Risk
31.31% 39.85%

Heart Attack Risk
31.31% 39.85%

Heart Attack Risk
27.12% 45.76%

Sex
☑ (All)
☑ Female
☑ Male

Measure Names
🟧 Avg. Smoking
🟩 Heart Attack..

Average Heart Attack Risk % by Countries given BMI

Line Chart: Average values of continuous variables affecting health versus age