In [152]:
```
import pandas as pd
import numpy as np
from sklearn.preprocessing import KBinsDiscretizer as kb
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_objs as go
```

Feeding the dataset into a Pandas data-frame

In [163]:
```
df = pd.read_csv('heart_attack_prediction_dataset.csv')
pd.set_option('display.max_columns', None)
df.head()
```

Out[163]:

	Patient ID	Age	Sex	Cholesterol	Blood Pressure	Heart Rate	Diabetes	Family History	Smoking	Obesity	Alcohol Consumption	Exercise Hours Per Week	Diet	Previous Heart Problems	Medicati U
0	BMW7812	67	Male	208	158/88	72	0	0	1	0	0	4.168189	Average	0	
1	CZE1114	21	Male	389	165/93	98	1	1	1	1	1	1.813242	Unhealthy	1	
2	BNI9906	21	Female	324	174/99	72	1	0	0	0	0	2.078353	Healthy	1	
3	JLN3497	84	Male	383	163/100	73	1	1	1	0	1	9.828130	Average	1	
4	GFO8847	66	Male	318	91/88	93	1	1	1	1	0	5.804299	Unhealthy	1	

Data Exploration

In [85]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8763 entries, 0 to 8762
Data columns (total 26 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   Patient ID                      8763 non-null   object 
 1   Age                             8763 non-null   int64  
 2   Sex                             8763 non-null   object 
 3   Cholesterol                     8763 non-null   int64  
 4   Blood Pressure                  8763 non-null   object 
 5   Heart Rate                      8763 non-null   int64  
 6   Diabetes                        8763 non-null   int64  
 7   Family History                  8763 non-null   int64  
 8   Smoking                         8763 non-null   int64  
 9   Obesity                         8763 non-null   int64  
 10  Alcohol Consumption             8763 non-null   int64  
 11  Exercise Hours Per Week         8763 non-null   float64
 12  Diet                            8763 non-null   object 
 13  Previous Heart Problems         8763 non-null   int64  
 14  Medication Use                  8763 non-null   int64  
 15  Stress Level                    8763 non-null   int64  
 16  Sedentary Hours Per Day         8763 non-null   float64
 17  Income                          8763 non-null   int64  
 18  BMI                             8763 non-null   float64
 19  Triglycerides                   8763 non-null   int64  
 20  Physical Activity Days Per Week 8763 non-null   int64  
 21  Sleep Hours Per Day             8763 non-null   int64  
 22  Country                         8763 non-null   object 
 23  Continent                       8763 non-null   object 
 24  Hemisphere                      8763 non-null   object 
 25  Heart Attack Risk               8763 non-null   int64  
dtypes: float64(3), int64(16), object(7)
memory usage: 1.7+ MB
```

In [162]:
```
df.describe()
```

Out[162]:

	Age	Cholesterol	Heart Rate	Diabetes	Family History	Smoking	Obesity	Alcohol Consumption	Exercise Hours Per Week	Previous Heart Problems	Medicat l
count	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000(
mean	53.707977	259.877211	75.021682	0.652288	0.492982	0.896839	0.501426	0.598083	10.014284	0.495835	0.498
std	21.249509	80.863276	20.550948	0.476271	0.499979	0.304186	0.500026	0.490313	5.783745	0.500011	0.500(
min	18.000000	120.000000	40.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.002442	0.000000	0.000(
25%	35.000000	192.000000	57.000000	0.000000	0.000000	1.000000	0.000000	0.000000	4.981579	0.000000	0.000(
50%	54.000000	259.000000	75.000000	1.000000	0.000000	1.000000	1.000000	1.000000	10.069559	0.000000	0.000(
75%	72.000000	330.000000	93.000000	1.000000	1.000000	1.000000	1.000000	1.000000	15.050018	1.000000	1.000(
max	90.000000	400.000000	110.000000	1.000000	1.000000	1.000000	1.000000	1.000000	19.998709	1.000000	1.000(

Data Cleaning

In [88]:
```
df.isnull().sum()
```

Out[88]:
```
Patient ID                        0
Age                               0
Sex                               0
Cholesterol                       0
Blood Pressure                    0
Heart Rate                        0
Diabetes                          0
Family History                    0
Smoking                           0
Obesity                           0
Alcohol Consumption               0
Exercise Hours Per Week           0
Diet                              0
Previous Heart Problems           0
Medication Use                    0
Stress Level                      0
Sedentary Hours Per Day           0
Income                            0
BMI                               0
Triglycerides                     0
Physical Activity Days Per Week   0
Sleep Hours Per Day               0
Country                           0
Continent                         0
Hemisphere                        0
Heart Attack Risk                 0
dtype: int64
```

In [89]:
```
df['Country'].unique()
```

Out[89]:
```
array(['Argentina', 'Canada', 'France', 'Thailand', 'Germany', 'Japan',
       'Brazil', 'South Africa', 'United States', 'Vietnam', 'China',
       'Italy', 'Spain', 'India', 'Nigeria', 'New Zealand', 'South Korea',
       'Australia', 'Colombia', 'United Kingdom'], dtype=object)
```

In [90]:
```
df['Country'].unique().size
```

Out[90]: 20

In [91]:
```
df.drop(columns=['Hemisphere'])
```

Female	324	174/99	72	1	0	0	0	...	9	9.463426	235282	28.176571	587	4	4	France	Europ
Male	383	163/100	73	1	1	1	0	...	9	7.648981	125640	36.464704	378	3	4	Canada	Nort Americ
Male	318	91/88	93	1	1	1	1	...	6	1.514821	160555	21.809144	231	1	5	Thailand	Asi
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
Male	121	94/76	61	1	1	1	0	...	8	10.806373	235420	19.655895	67	7	7	Thailand	Asi
Female	120	157/102	73	1	0	0	1	...	8	3.833038	217881	23.993866	617	4	9	Canada	Nort Americ
Male	250	161/75	105	0	1	1	1	...	5	2.375214	36998	35.406146	527	4	4	Brazil	Sout Americ
Male	178	119/67	60	1	0	1	0	...	5	0.029104	209943	27.294020	114	2	8	Brazil	Sout Americ
Female	356	138/67	75	1	1	0	0	...	8	9.005234	247338	32.914151	180	7	4	United Kingdom	Europ

s

Data Transformation

Discretization

In [92]:
```
num_col = ['Age','Income']
num_data = df[num_col]
disc = kb(n_bins=10, encode='ordinal', strategy='uniform')
data_disc = disc.fit_transform(num_data)
for i, col in enumerate(num_col):
    df[col] = data_disc[:, i]
df.head()
```

Out[92]:

	Patient ID	Age	Sex	Cholesterol	Blood Pressure	Heart Rate	Diabetes	Family History	Smoking	Obesity	...	Sedentary Hours Per Day	Income	BMI	Triglycerides	Physi Activ Da P We
0	BMW7812	6.0	Male	208	158/88	72	0	0	1	0	...	6.615001	8.0	31.251233	286	
1	CZE1114	0.0	Male	389	165/93	98	1	1	1	1	...	4.963459	9.0	27.194973	235	
2	BNI9906	0.0	Female	324	174/99	72	1	0	0	0	...	9.463426	7.0	28.176571	587	
3	JLN3497	9.0	Male	383	163/100	73	1	1	1	0	...	7.648981	3.0	36.464704	378	
4	GFO8847	6.0	Male	318	91/88	93	1	1	1	1	...	1.514821	5.0	21.809144	231	

5 rows × 26 columns

Data Filtering

In [120]:
```
count = df['Sex'].value_counts().reset_index()
count.columns = ['Sex','Count']
count
```

Out[120]:

	Sex	Count
0	Male	6111
1	Female	2652

In [103]:
```
new = df[['Sex', 'Heart Attack Risk']].groupby('Sex').sum().reset_index()
new
```

Out[103]:

	Sex	Heart Attack Risk
0	Female	944
1	Male	2195

In [104]:
```
new_df = pd.merge(new, count, on='Sex')
new_df
```

Out[104]:

	Sex	Heart Attack Risk	Count
0	Female	944	2652
1	Male	2195	6111

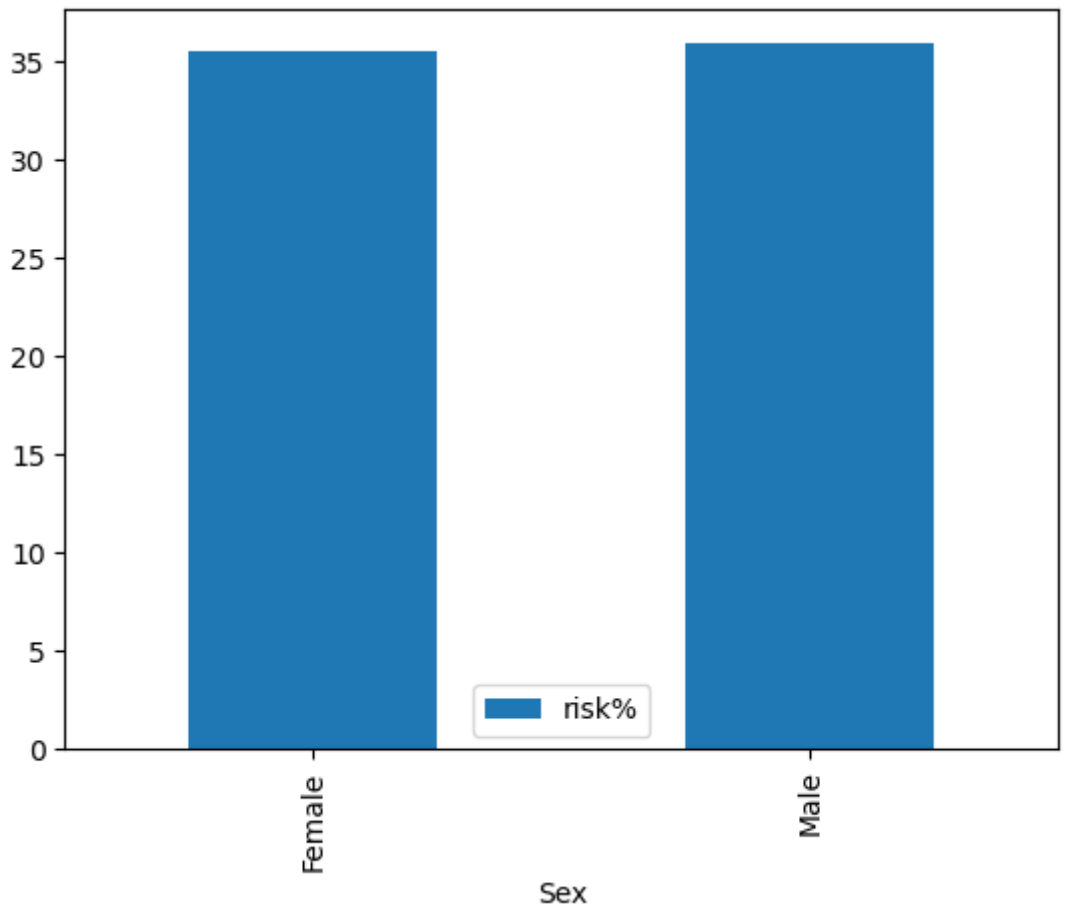
In [109]:
```
new_df['risk%'] = (new_df['Heart Attack Risk']/new_df['Count'])*100
new_df
```

Out[109]:

	Sex	Heart Attack Risk	Count	risk%
0	Female	944	2652	35.595777
1	Male	2195	6111	35.918835

In [111]:
```
new_df.plot.bar('Sex','risk%')
```

Out[111]: <Axes: xlabel='Sex'>

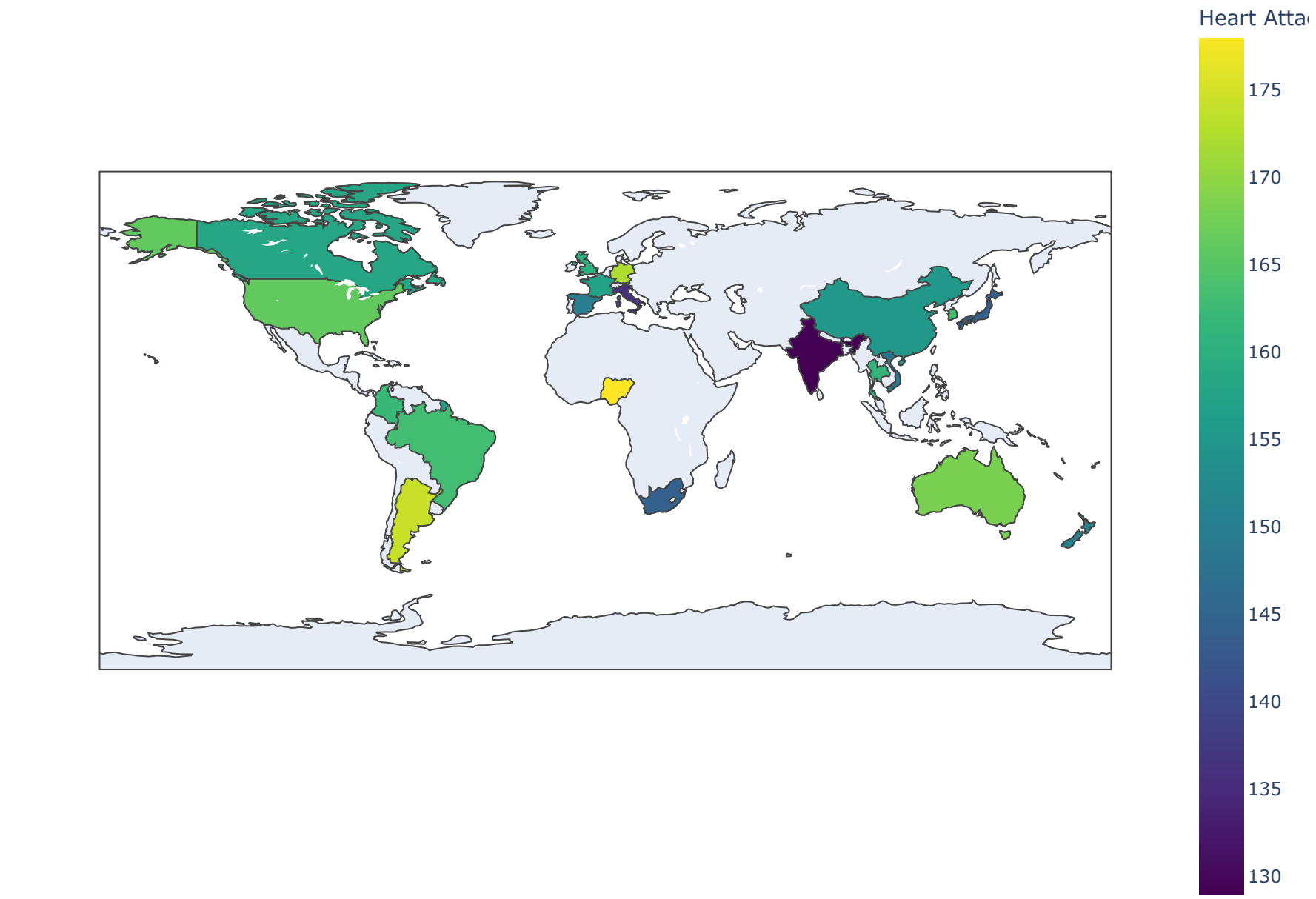


In [150]:
```
con_risk = df[['Country' , 'Heart Attack Risk']] .groupby('Country').sum().reset_index()
con_risk
```

Out[150]:

	Country	Heart Attack Risk
0	Argentina	174
1	Australia	168
2	Brazil	163
3	Canada	158
4	China	155
5	Colombia	162
6	France	157
7	Germany	172
8	India	129
9	Italy	136
10	Japan	144
11	New Zealand	151
12	Nigeria	178
13	South Africa	144
14	South Korea	163
15	Spain	150
16	Thailand	161
17	United Kingdom	160
18	United States	166
19	Vietnam	148

In [158]:
```
fig = px.choropleth(con_risk,
                    locations='Country',
                    locationmode='country names',
                    color='Heart Attack Risk',
                    color_continuous_scale='Viridis',
                    width=1000,
                    height=800,
                    title='Heart Attack Risk by Country')
fig.show()
```

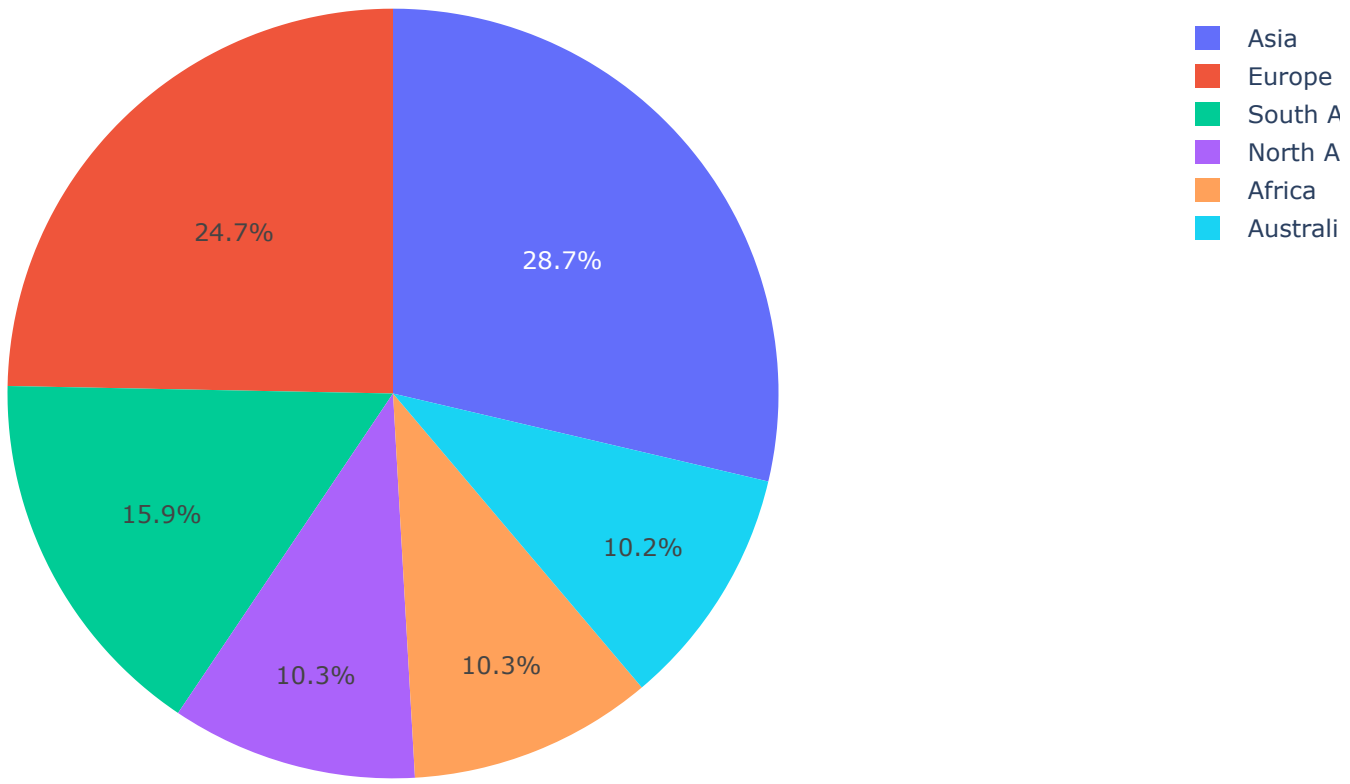


In [154]:
```
conti_risk = df[['Continent' , 'Heart Attack Risk']] .groupby('Continent').sum().reset_index()
conti_risk
```

Out[154]:

	Continent	Heart Attack Risk
0	Africa	322
1	Asia	900
2	Australia	319
3	Europe	775
4	North America	324
5	South America	499

In [155]:
```
fig = px.pie(conti_risk , values='Heart Attack Risk' , names='Continent')
fig.show()
```



In [160]:
```
lifestyle_risk = df.groupby('Heart Attack Risk')[['Smoking', 'Previous Heart Problems', 'Obesity', 'Alcohol Cons
ax = lifestyle_risk.T.plot(kind='bar', stacked=True, colormap='Set3', figsize=(10, 6))
ax.set_xlabel('Lifestyle Factors')
ax.set_ylabel('Risk')
ax.set_title('Relationship Between Lifestyle Factors and Heart Attack Risk')
ax.legend(title='Heart Attack Risk', labels=['No Risk (0)', 'Risk (1)'], loc='upper left')

plt.xticks(rotation=0)
plt.tight_layout()
plt.show()
```

In [164]:
```
col_of_int = ['Age', 'BMI', 'Cholesterol', 'Heart Attack Risk','Income']
df_subset = df[col_of_int]
cor_matrix = df_subset.corr()
plt.figure(figsize=(8, 6))
sns.heatmap(cor_matrix, annot=True, cmap='coolwarm', fmt='.2f', linewidths=0.5)
plt.title('Correlation Heatmap')
plt.show()
```