Pandas and Seaborn based homework

DSE5002

Beija Richardson 3/30/25

We will be working with the heart.csv data set

https://www.kaggle.com/fedesoriano/heart-failure-prediction?select=heart.csv

using tools in pandas and seaborn, and ideas from the two Jupyter notebooks we've seen this week

In [101…

```
!pip install seaborn
import pandas as pd
import numpy as np
import seaborn as sns
import p9
```

```
Requirement already satisfied: seaborn in c:\users\luke\anaconda3\envs\class5002\lib
\site-packages (0.13.2)
Requirement already satisfied: numpy!=1.24.0,>=1.20 in c:\users\luke\anaconda3\envs
\class5002\lib\site-packages (from seaborn) (2.0.1)
Requirement already satisfied: pandas>=1.2 in c:\users\luke\anaconda3\envs\class5002
\lib\site-packages (from seaborn) (2.2.3)
Requirement already satisfied: matplotlib!=3.6.1,>=3.4 in c:\users\luke\anaconda3\en
vs\class5002\lib\site-packages (from seaborn) (3.10.1)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\luke\anaconda3\envs\clas
s5002\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (1.3.1)
Requirement already satisfied: cycler>=0.10 in c:\users\luke\anaconda3\envs\class500
2\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\luke\anaconda3\envs\cla
ss5002\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (4.56.0)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\luke\anaconda3\envs\cla
ss5002\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (1.4.8)
Requirement already satisfied: packaging>=20.0 in c:\users\luke\anaconda3\envs\class
5002\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (24.2)
Requirement already satisfied: pillow>=8 in c:\users\luke\anaconda3\envs\class5002\l
ib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (11.1.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\luke\anaconda3\envs\clas
s5002\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (3.2.3)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\luke\anaconda3\envs
\class5002\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in c:\users\luke\anaconda3\envs\class500
2\lib\site-packages (from pandas>=1.2->seaborn) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in c:\users\luke\anaconda3\envs\class5
002\lib\site-packages (from pandas>=1.2->seaborn) (2023.3)
Requirement already satisfied: six>=1.5 in c:\users\luke\anaconda3\envs\class5002\li
b\site-packages (from python-dateutil>=2.7->matplotlib!=3.6.1,>=3.4->seaborn) (1.16.
0)
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
Cell In[101], line 5
      3 import numpy as np
      4 import seaborn as sns
----> 5 import p9

ModuleNotFoundError: No module named 'p9'
```

In [39]:

```
# make sure heart.csv is in your current working directory, or list the full path n

infile="C:/Users/Luke/Documents/Class5002/Module_2/Practice Exercises\\heart.csv"

bp_df=pd.read_csv(r"C:/Users/Luke/Documents/Class5002/Module_2/Practice Exercises/h
bp_df.head()
```

Out[39]:

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	Exer
0	40	M	ATA	140	289	0	Normal	172	
1	49	F	NAP	160	180	0	Normal	156	
2	37	M	ATA	130	283	0	ST	98	
3	48	F	ASY	138	214	0	Normal	108	
4	54	M	NAP	150	195	0	Normal	122	

Find or create the following

a.) -Find the dimensions, memory used, and other basic information

b.) -Run the data summary

c.) Change the appropriate variables to type Categorical

d.) -Create a pivot table (using the Pandas groupby operation) showing mean Resting BP by Sex, Resting ECG and HeartDisease-What does this tell you? What else can you figure out using a Pivot table, show me two other helpful pivot tables based on different variables, different groupings or different aggregation functions (count, mean, max etc)

e.) -Show a histogram and the ECDF (empirical cumulative distribution function) for several continuous variables in the data set, in broad terms, what do the distributions look like, normal? exponential, poison-like?, uniform? Does this match your expectations?

```
https://seaborn.pydata.org/generated/seaborn.ecdfplot.html
https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.ecdf.html
```

f.) -Show An SNS Pairplot, the most informative version you can find, set the hue based on Heart Disease, try using at least one other variable as the Hue. Discuss what you think you are seeing in this plot

Create all these results in this Notebook and turn it in

Responses

```
In [41]: print(bp_df.shape)
```

```
(918, 12)
```

```
In [43]: bp_df.memory_usage(deep=True)
```

```
Out[43]: Index              132
         Age               7344
         Sex              45900
         ChestPainType    47690
         RestingBP         7344
         Cholesterol       7344
         FastingBS         7344
         RestingECG       49214
         MaxHR             7344
         ExerciseAngina   45900
         Oldpeak           7344
         ST_Slope         47864
         HeartDisease      7344
         dtype: int64
```

```
In [45]: bp_df.describe()
```

Out[45]:

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisea
count	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.0000
mean	53.510893	132.396514	198.799564	0.233115	136.809368	0.887364	0.5533
std	9.432617	18.514154	109.384145	0.423046	25.460334	1.066570	0.4974
min	28.000000	0.000000	0.000000	0.000000	60.000000	-2.600000	0.0000
25%	47.000000	120.000000	173.250000	0.000000	120.000000	0.000000	0.0000
50%	54.000000	130.000000	223.000000	0.000000	138.000000	0.600000	1.0000
75%	60.000000	140.000000	267.000000	0.000000	156.000000	1.500000	1.0000
max	77.000000	200.000000	603.000000	1.000000	202.000000	6.200000	1.0000

```
In [47]: print(bp_df.nunique())
```

```
Age                50
Sex                 2
ChestPainType       4
RestingBP          67
Cholesterol       222
FastingBS           2
RestingECG          3
MaxHR             119
ExerciseAngina      2
Oldpeak            53
ST_Slope            3
HeartDisease        2
dtype: int64
```

In [49]:
```
pivot_table_1 = bp_df.groupby(['Sex', 'RestingECG', 'HeartDisease'])['RestingBP'].m
```

In [51]:
```
print(pivot_table_1)
```

```
   Sex RestingECG  HeartDisease   RestingBP
0    F        LVH             0  128.696970
1    F        LVH             1  148.928571
2    F     Normal             0  129.123596
3    F     Normal             1  139.310345
4    F         ST             0  127.523810
5    F         ST             1  139.285714
6    M        LVH             0  131.836735
7    M        LVH             1  135.467391
8    M     Normal             0  129.921348
9    M     Normal             1  130.675781
10   M         ST             0  134.275000
11   M         ST             1  137.727273
```

In [53]:
```
continuous_vars = ['Age', 'RestingBP', 'Cholesterol', 'MaxHR']
```

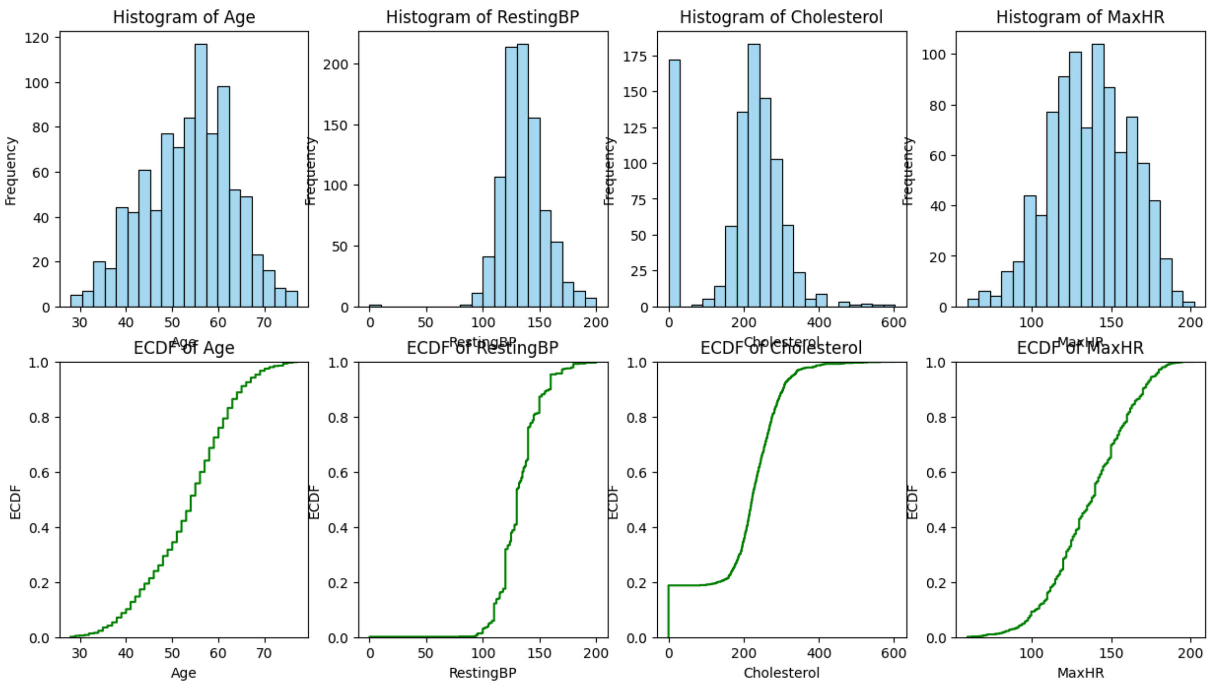
In [87]:
```
pip install matplotlib seaborn numpy
```

```
Requirement already satisfied: matplotlib in c:\users\luke\anaconda3\envs\class5002
\lib\site-packages (3.10.1)Note: you may need to restart the kernel to use updated p
ackages.

Requirement already satisfied: seaborn in c:\users\luke\anaconda3\envs\class5002\lib
\site-packages (0.13.2)
Requirement already satisfied: numpy in c:\users\luke\anaconda3\envs\class5002\lib\s
ite-packages (2.0.1)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\luke\anaconda3\envs\clas
s5002\lib\site-packages (from matplotlib) (1.3.1)
Requirement already satisfied: cycler>=0.10 in c:\users\luke\anaconda3\envs\class500
2\lib\site-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\luke\anaconda3\envs\cla
ss5002\lib\site-packages (from matplotlib) (4.56.0)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\luke\anaconda3\envs\cla
ss5002\lib\site-packages (from matplotlib) (1.4.8)
Requirement already satisfied: packaging>=20.0 in c:\users\luke\anaconda3\envs\class
5002\lib\site-packages (from matplotlib) (24.2)
Requirement already satisfied: pillow>=8 in c:\users\luke\anaconda3\envs\class5002\l
ib\site-packages (from matplotlib) (11.1.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\luke\anaconda3\envs\clas
s5002\lib\site-packages (from matplotlib) (3.2.3)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\luke\anaconda3\envs
\class5002\lib\site-packages (from matplotlib) (2.9.0.post0)
Requirement already satisfied: pandas>=1.2 in c:\users\luke\anaconda3\envs\class5002
\lib\site-packages (from seaborn) (2.2.3)
Requirement already satisfied: pytz>=2020.1 in c:\users\luke\anaconda3\envs\class500
2\lib\site-packages (from pandas>=1.2->seaborn) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in c:\users\luke\anaconda3\envs\class5
002\lib\site-packages (from pandas>=1.2->seaborn) (2023.3)
Requirement already satisfied: six>=1.5 in c:\users\luke\anaconda3\envs\class5002\li
b\site-packages (from python-dateutil>=2.7->matplotlib) (1.16.0)
```

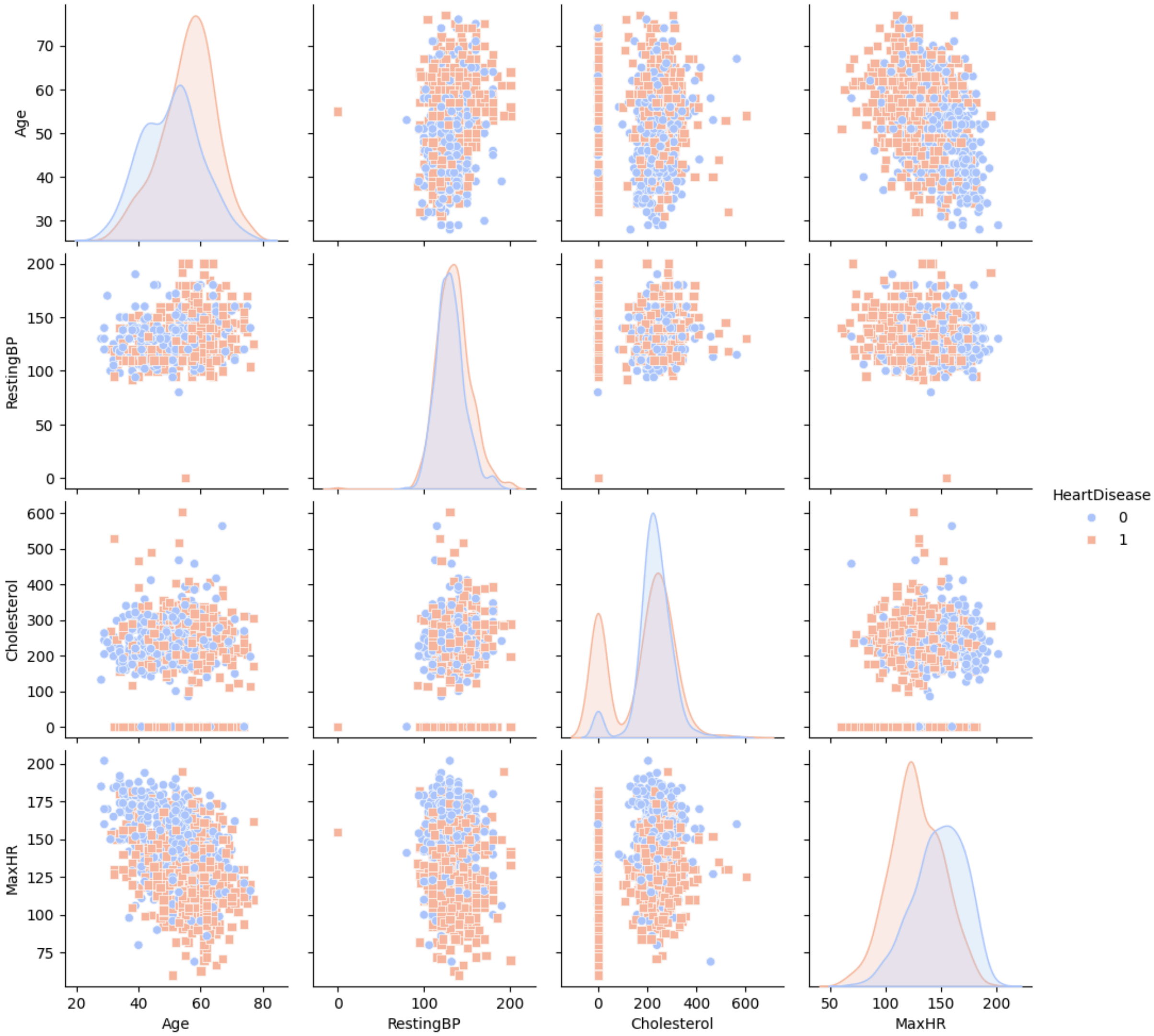
In [91]:

```python
import matplotlib.pyplot as plt
fig, axes = plt.subplots(2, len(continuous_vars), figsize=(15, 8))
continuous_vars = ['Age', 'RestingBP', 'Cholesterol', 'MaxHR']
for i, var in enumerate(continuous_vars):
    sns.histplot(bp_df[var], kde=False, ax=axes[0, i], bins=20, color='skyblue')
    axes[0, i].set_title(f'Histogram of {var}')
    axes[0, i].set_xlabel(var)
    axes[0, i].set_ylabel('Frequency')
    sns.ecdfplot(bp_df[var], ax=axes[1, i], color='green')
    axes[1, i].set_title(f'ECDF of {var}')
    axes[1, i].set_xlabel(var)
    axes[1, i].set_ylabel('ECDF')
```



In [93]:
```
sns.pairplot(bp_df[continuous_vars + ['HeartDisease', 'Sex']], hue='HeartDisease',
```

Out[93]:
```
<seaborn.axisgrid.PairGrid at 0x131c4f12030>
```



```
In [97]: print(dp_df)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[97], line 1
----> 1 print(dp_df)

NameError: name 'dp_df' is not defined
```

```
In [ ]: g.) Create several useful or informative boxplots of continuous variables by catego
        among the variables,   discuss what you think it means or implies

        h.) Create violin plots of these same results
```

```
In [103… sns.boxplot(x='HeartDisease', y='RestingBP', data=bp_df, ax=axes[0, 1], palette='co
         axes[0, 1].set_title('Resting BP by Heart Disease')
         axes[0, 1].set_xlabel('Heart Disease')
         axes[0, 1].set_ylabel('Resting BP')
```

```
C:\Users\Luke\AppData\Local\Temp\ipykernel_26632\1096069617.py:1: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.1
4.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.boxplot(x='HeartDisease', y='RestingBP', data=bp_df, ax=axes[0, 1], palette='c
oolwarm')
```

Out[103…

```
Text(327.2608695652175, 0.5, 'Resting BP')
```

In [ ]:

```
1.) Find the mean, median and standard deviation of the Max heartrate variable in t

Turn this into a pivot table,   grouping by one or more predictors.
```

In [113…

```
maxhr_mean = bp_df['MaxHR'].mean()
maxhr_median = bp_df['MaxHR'].median()
maxhr_std = bp_df['MaxHR'].std()
print(f"Mean of MaxHR: {maxhr_mean}")
print(f"Median of MaxHR: {maxhr_median}")
print(f"Standard Deviation of MaxHR: {maxhr_std}")
```

```
Mean of MaxHR: 136.80936819172112
Median of MaxHR: 138.0
Standard Deviation of MaxHR: 25.460334138250293
```

In [115…

```
pivot_table = bp_df.groupby(['HeartDisease', 'Sex'])['MaxHR'].agg(['mean', 'median'
```

In [117…

```
print(pivot_table)
```

```
                        mean  median        std
HeartDisease Sex                               
0            F    149.048951   152.0  21.597903
             M    147.670412   150.0  24.170369
1            F    137.820000   142.5  21.820876
             M    126.545852   125.0  23.306611
```

In [ ]: