Brian Reppeto DSC630 Week 1_2

September 1, 2024

DSC 630 Week 1 :

Activity 1.2

Author: Brian Reppeto 8/25/2024

[1]:
```
# import libraries

import pandas as pd
```

[2]:
```
# import heart disease data set

heart_df=pd.read_csv("heart_disease_uci.csv")
```

[3]:
```
# display the first 10 rows

heart_df.head(10)
```

[3]:

	id	age	sex	dataset	cp	trestbps	chol	fbs	restecg	thalch	exang	oldpeak	slope	ca
0	1	63	Male	Cleveland	typical angina	145.0	233.0	True	lv hypertrophy	150.0	False	2.3	downsloping	0.0
1	2	67	Male	Cleveland	asymptomatic	160.0	286.0	False	lv hypertrophy	108.0	True	1.5	flat	3.0
2	3	67	Male	Cleveland	asymptomatic	120.0	229.0	False	lv hypertrophy	129.0	True	2.6	flat	2.0
3	4	37	Male	Cleveland	non-anginal	130.0	250.0	False	normal	187.0	False	3.5	downsloping	0.0
4	5	41	Female	Cleveland	atypical angina	130.0	204.0	False	lv hypertrophy	172.0	False	1.4	upsloping	0.0
5	6	56	Male	Cleveland	atypical angina	120.0	236.0	False	normal	178.0	False	0.8	upsloping	0.0
6	7	62	Female	Cleveland	asymptomatic	140.0	268.0	False	lv hypertrophy	160.0	False	3.6	downsloping	2.0
7	8	57	Female	Cleveland	asymptomatic	120.0	354.0	False	normal	163.0	True	0.6	upsloping	0.0
8	9	63	Male	Cleveland	asymptomatic	130.0	254.0	False						
9	10	53	Male	Cleveland	asymptomatic	140.0	203.0	True						

```
8  lv hypertrophy   147.0  False      1.4         flat  1.0
9  lv hypertrophy   155.0   True      3.1  downsloping  0.0

                thal  num
0       fixed defect    0
1             normal    2
2  reversable defect    1
3             normal    0
4             normal    0
5             normal    0
6             normal    3
7             normal    0
8  reversable defect    2
9  reversable defect    1
```

[4]:
```
# summary statistics of the dataset to understand its structure and content

df_summary = heart_df.describe(include='all')

df_summary
```

[4]:
```
                id         age   sex    dataset            cp    trestbps  \
count   920.000000  920.000000   920        920           920  861.000000
unique         NaN         NaN     2          4             4         NaN
top            NaN         NaN  Male  Cleveland  asymptomatic         NaN
freq           NaN         NaN   726        304           496         NaN
mean    460.500000   53.510870   NaN        NaN           NaN  132.132404
std     265.725422    9.424685   NaN        NaN           NaN   19.066070
min       1.000000   28.000000   NaN        NaN           NaN    0.000000
25%     230.750000   47.000000   NaN        NaN           NaN  120.000000
50%     460.500000   54.000000   NaN        NaN           NaN  130.000000
75%     690.250000   60.000000   NaN        NaN           NaN  140.000000
max     920.000000   77.000000   NaN        NaN           NaN  200.000000

              chol    fbs restecg      thalch  exang     oldpeak slope  \
count   890.000000    830     918  865.000000    865  858.000000   611
unique         NaN      2       3         NaN      2         NaN     3
top            NaN  False  normal         NaN  False         NaN  flat
freq           NaN    692     551         NaN    528         NaN   345
mean    199.130337    NaN     NaN  137.545665    NaN    0.878788   NaN
std     110.780810    NaN     NaN   25.926276    NaN    1.091226   NaN
min       0.000000    NaN     NaN   60.000000    NaN   -2.600000   NaN
25%     175.000000    NaN     NaN  120.000000    NaN    0.000000   NaN
50%     223.000000    NaN     NaN  140.000000    NaN    0.500000   NaN
75%     268.000000    NaN     NaN  157.000000    NaN    1.500000   NaN
max     603.000000    NaN     NaN  202.000000    NaN    6.200000   NaN
```

	ca	thal	num
count	309.000000	434	920.000000
unique	NaN	3	NaN
top	NaN	normal	NaN
freq	NaN	196	NaN
mean	0.676375	NaN	0.995652
std	0.935653	NaN	1.142693
min	0.000000	NaN	0.000000
25%	0.000000	NaN	0.000000
50%	0.000000	NaN	1.000000
75%	1.000000	NaN	2.000000
max	3.000000	NaN	4.000000

1 Data Summary

The dataset contains various medical attributes used to diagnose heart disease, and includes the following columns:

- id: Identifier for the record.
- age: Age of the patient.
- sex: Gender of the patient (Male/Female).
- dataset: The dataset origin (e.g., Cleveland).
- cp: Type of chest pain experienced.
- trestbps: Resting blood pressure (in mm Hg).
- chol: Serum cholesterol (in mg/dl).
- fbs: Fasting blood sugar > 120 mg/dl (True/False).
- restecg: Resting electrocardiographic results.
- thalch: Maximum heart rate achieved.
- exang: Exercise-induced angina (True/False).
- oldpeak: ST depression induced by exercise relative to rest.
- slope: Slope of the peak exercise ST segment.
- ca: Number of major vessels colored by fluoroscopy.
- thal: Thalassemia (a blood disorder).
- num: Diagnosis of heart disease (values 0-4).

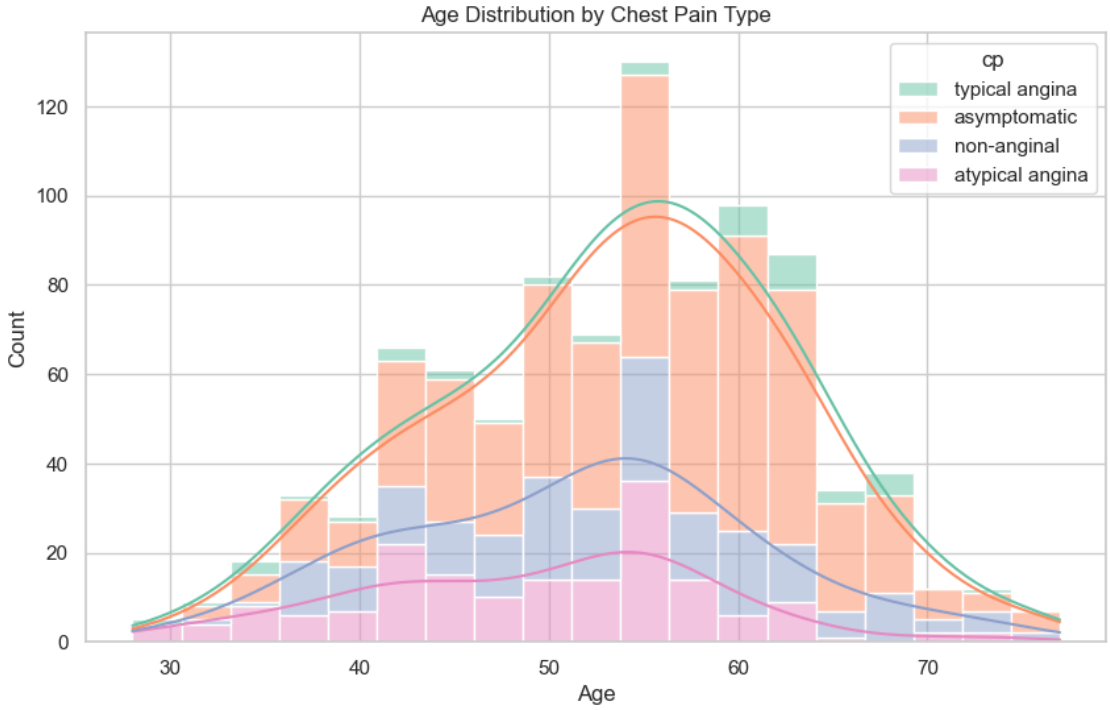
2 Questions to Explore

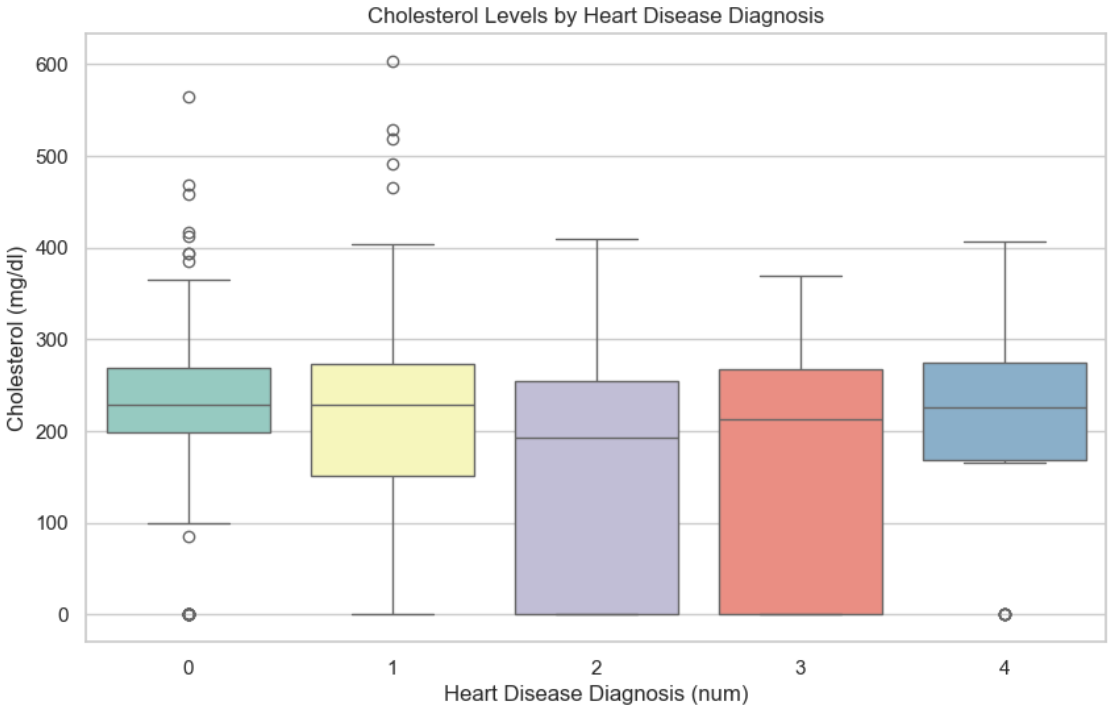
1. How does age distribution vary among patients with different types of chest pain?
2. Is there a significant difference in cholesterol levels between patients with and without heart disease?

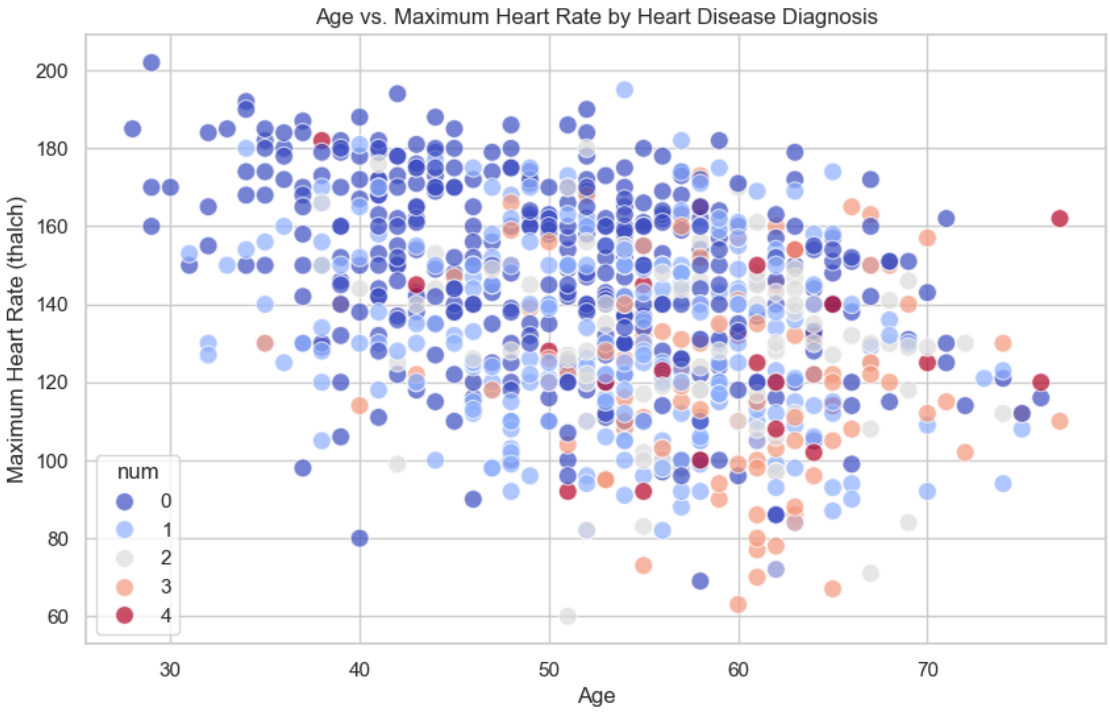
3 Visualizations

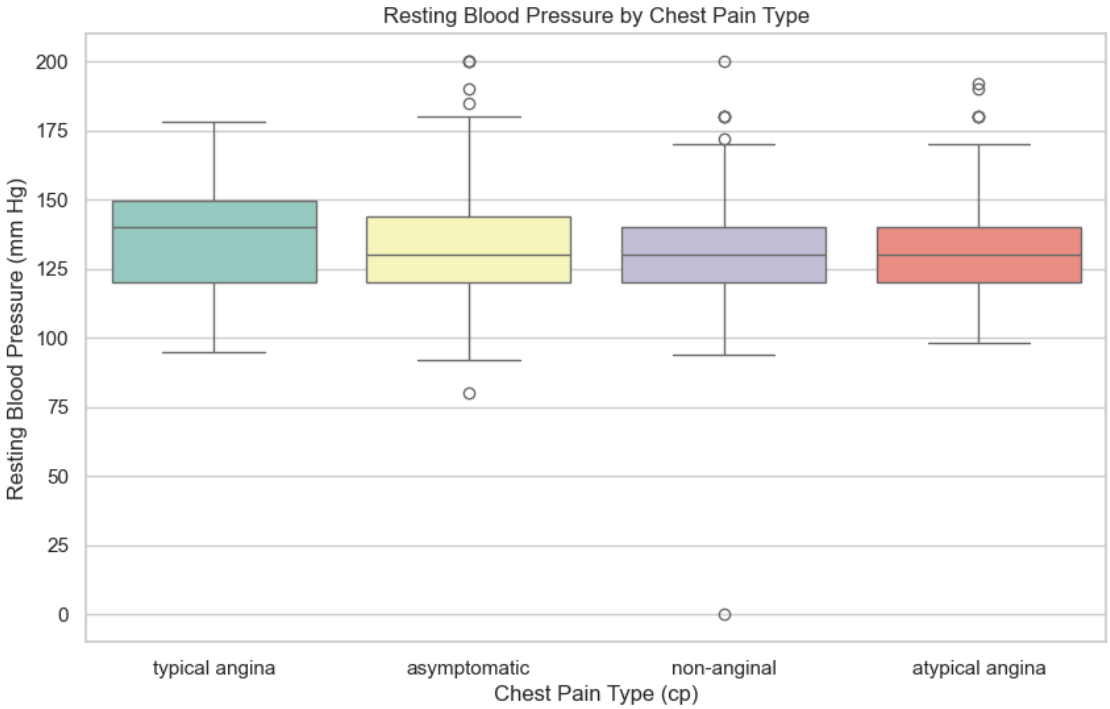
[5]:
```
# import libraries
```

```
import matplotlib.pyplot as plt
import seaborn as sns

# aesthetics for the plots
sns.set(style="whitegrid")

# histogram for age distribution by chest pain type
plt.figure(figsize=(10, 6))
sns.histplot(data=heart_df, x='age', hue='cp', multiple='stack', kde=True,␣
 ↪palette='Set2')
plt.title('Age Distribution by Chest Pain Type')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()

# boxplot to compare cholesterol levels by heart disease diagnosis
plt.figure(figsize=(10, 6))
sns.boxplot(data=heart_df, x='num', y='chol', hue='num', palette='Set3',␣
 ↪legend=False)
plt.title('Cholesterol Levels by Heart Disease Diagnosis')
plt.xlabel('Heart Disease Diagnosis (num)')
plt.ylabel('Cholesterol (mg/dl)')
plt.show()

# bivariate plot to explore the relationship between age and maximum heart rate
plt.figure(figsize=(10, 6))
sns.scatterplot(data=heart_df, x='age', y='thalch', hue='num',␣
 ↪palette='coolwarm', s=100, alpha=0.7)
plt.title('Age vs. Maximum Heart Rate by Heart Disease Diagnosis')
plt.xlabel('Age')
plt.ylabel('Maximum Heart Rate (thalch)')
plt.show()

# additional visualization: boxplot for resting blood pressure by chest pain␣
 ↪type
plt.figure(figsize=(10, 6))
sns.boxplot(data=heart_df, x='cp', y='trestbps', hue='cp', palette='Set3',␣
 ↪legend=False)
plt.title('Resting Blood Pressure by Chest Pain Type')
plt.xlabel('Chest Pain Type (cp)')
plt.ylabel('Resting Blood Pressure (mm Hg)')
plt.show()
```


Age Distribution by Chest Pain Type
cp
typical angina
asymptomatic
non-anginal
atypical angina
Count
Age
0
20
40
60
80
100
120
30
40
50
60
70


Cholesterol Levels by Heart Disease Diagnosis
Cholesterol (mg/dl)
Heart Disease Diagnosis (num)
0
100
200
300
400
500
600
0
1
2
3
4


Age vs. Maximum Heart Rate by Heart Disease Diagnosis
Maximum Heart Rate (thalch)
Age
num
0
1
2
3
4
200
180
160
140
120
100
80
60
30
40
50
60
70


Resting Blood Pressure by Chest Pain Type
Resting Blood Pressure (mm Hg)
Chest Pain Type (cp)
200
175
150
125
100
75
50
25
0
typical angina
asymptomatic
non-anginal
atypical angina

4 Summary of Results

1. Age Distribution by Chest Pain Type: The histogram reveals that age distribution varies across different types of chest pain. Patients with asymptomatic chest pain tend to be older, whereas those with atypical angina are generally younger. This suggests that chest pain types might correlate with age in the context of heart disease.
2. Cholesterol Levels by Heart Disease Diagnosis: The boxplot shows that cholesterol levels are generally higher in patients with heart disease (num > 0) compared to those without (num $= 0$). However, there is a significant overlap, indicating that while cholesterol level is a factor, it is not the sole determinant of heart disease.
3. Age vs. Maximum Heart Rate by Heart Disease Diagnosis: The scatter plot suggests a weak negative correlation between age and maximum heart rate, especially in patients with heart disease (num > 0). Older patients tend to have lower maximum heart rates, and those with heart disease often show reduced heart rate capacity.
4. Resting Blood Pressure by Chest Pain Type: The additional boxplot indicates that resting blood pressure varies slightly by chest pain type, with typical and asymptomatic angina patients showing slightly higher resting blood pressure. However, the differences are not very pronounced.

5 Conclusion

The visualizations support the conclusion that age, cholesterol levels, and maximum heart rate are important factors in diagnosing heart disease. Patients with asymptomatic chest pain are generally older and have higher cholesterol levels, indicating a potential link to heart disease. The weak correlation between age and maximum heart rate also suggests that reduced heart rate capacity in older individuals might be a marker for heart disease. However, these factors alone do not provide a definitive diagnosis, highlighting the multifaceted nature of heart disease.

5.1 Dataset Citation

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988). *Heart Disease*. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X

[ ]: