心脏病跟什么有关？数据分析帮你看看

众所周知，心脏病已经成为人类健康的三大杀手之一。

随着父母岁数慢慢的变大了，身体也容易不好，身边的老人们在体检的时候，很多都发现了心脏不太好的情况。

因此逗汁儿就去 KAGGLE 上找了个心脏病的数据集，做探索性数据分析，简单的研究一下心脏病与哪些因素有关。

初步观察数据

这个数据集很小，仅有 303 行。我们来看一下数据吧~

首先导入相关的库，因为后续还要作图进行分析，我们就按照惯例给肯定需要用的库包括绘图库都导入。

```
import pandas  as  pd
import numpy  as  np
import matplotlib.pyplot  as  plt
import seaborn  as  sns
```

为了让 jupyter 里面显示出来图，千万不要忘记加这一行代码：

```
%matplotlib inline
```

读取数据，看下前十行：

```
df  =pd.read_csv("E:\\heart.csv")
df.head(10)
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

这里一共显示了 14 个字段，分别看一下这些字段代表的意思。

age:代表患者年龄，以年为单位。
sex:代表患者性别，1 为男性，0 为女性。
cp:胸痛类型，具体类型未说明
trestbps：入院前的静息血压，单位为毫米汞柱
chol:患者血清胆固醇含量，单位 mg/dl
fbs:患者空腹血糖大于 120mg/dl，1 为是，0 为否
restecg：静态心电图结果
thalach:达到最大的心率值
exang：运动诱发心绞痛，1 为是，0 为否。
oldspeak：相对于休息来说运动引起的 ST 段抑制
slope: 运动引起的 ST 段最高值斜率
ca: 通过荧光检测技术显示出来的主要血管数量（0-3）
thal:1 为正常，2 为固定缺陷，3 为可逆缺陷。
target：0 或 1 为是否患病

下面就看一下数据的主要信息

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
age         303 non-null int64
sex         303 non-null int64
cp          303 non-null int64
trestbps    303 non-null int64
chol        303 non-null int64
fbs         303 non-null int64
restecg     303 non-null int64
thalach     303 non-null int64
exang       303 non-null int64
oldpeak     303 non-null float64
slope       303 non-null int64
ca          303 non-null int64
thal        303 non-null int64
target      303 non-null int64
dtypes: float64(1), int64(13)
memory usage: 33.2 KB
```

可以看到每个字段的数据都非常完整，没有缺失的字段，数据类型也主要以整数型为主，其中一个是浮点型。

在这组数据中，有年龄等相关指标，我们来看看这些指标的平均值，最大最小值等常规统计学数据。

```
df.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

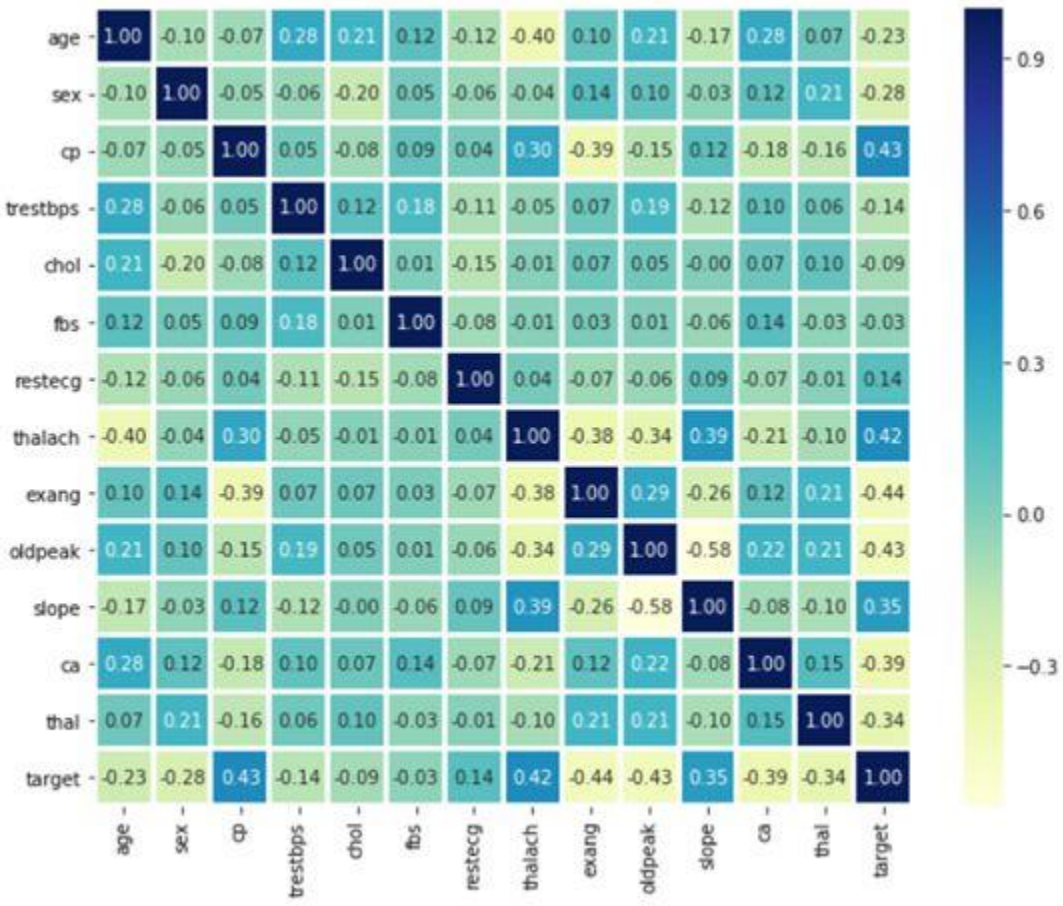
我们可以看到，平均年龄在 54 岁左右，最小的年龄在 29 岁，最大的年龄在 77 岁。因此这个数据集中，主要选取的是中老年人。

count 数值代表并无空值，这个上面分析的时候也已经看到了~。下面我们就对这些数据进行一下可视化分析。

相关数据的可视化分析

首先看下热图，整体看一下相关性

```
plt.figure(figsize=(10,8))
sns.heatmap(df.corr(),annot=True,cmap='YlGnBu',fmt='.2f',linewidths=2)
```

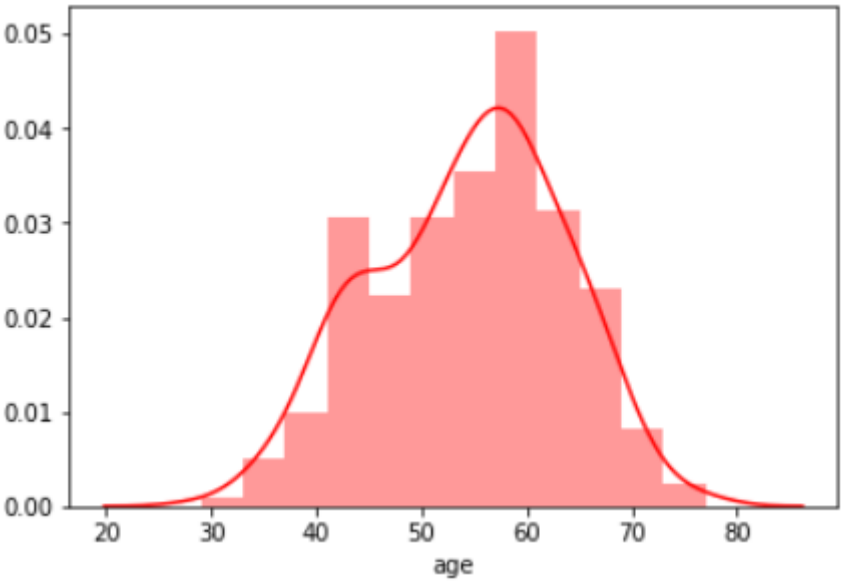


从热图上来看，只有 cp,thalach,slope 这三个指标与患病的相关性相对偏高，但是并不大。

看下患病人数有多少人，整个数据集中年龄集中在哪个段呢？

```
df['target'].value_counts()
sns.distplot(df['age'],color='Red')
```

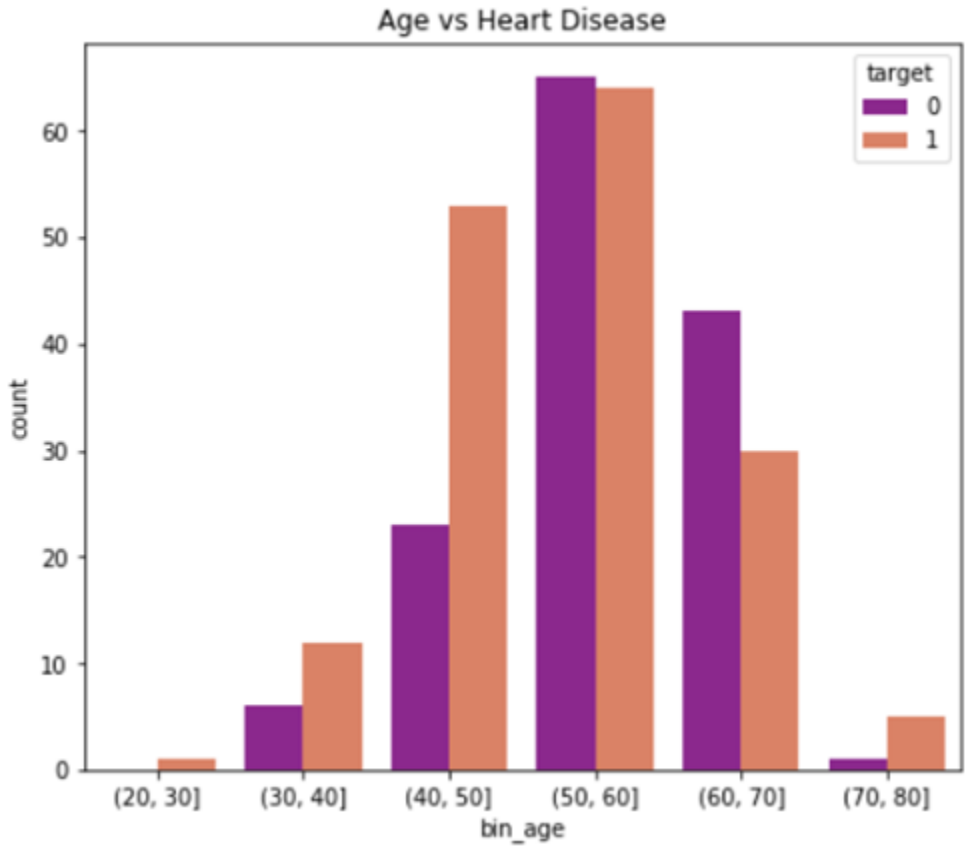
303 个数据中，有 165 个患病者...比例超过了一半。



大多数的人年龄都在 40-70 岁左右，偏向于中老年。

01　年龄与心脏病

```
fig,ax=plt.subplots(figsize=(24,6))
plt.subplot(1,3,  1)
age_bins= [20,30,40,50,60,70,80]
df['bin_age']=pd.cut(df['age'],bins=age_bins)
g1=sns.countplot(x='bin_age',data=df,hue='target',palette='plasma')
g1.set_title("Agevs Heart Disease")
```



大多数的人在 45-65 左右会经常去进行心脏病的检查，看看是否患有心脏病，也说明大家对于心脏病的还是比较恐惧的，

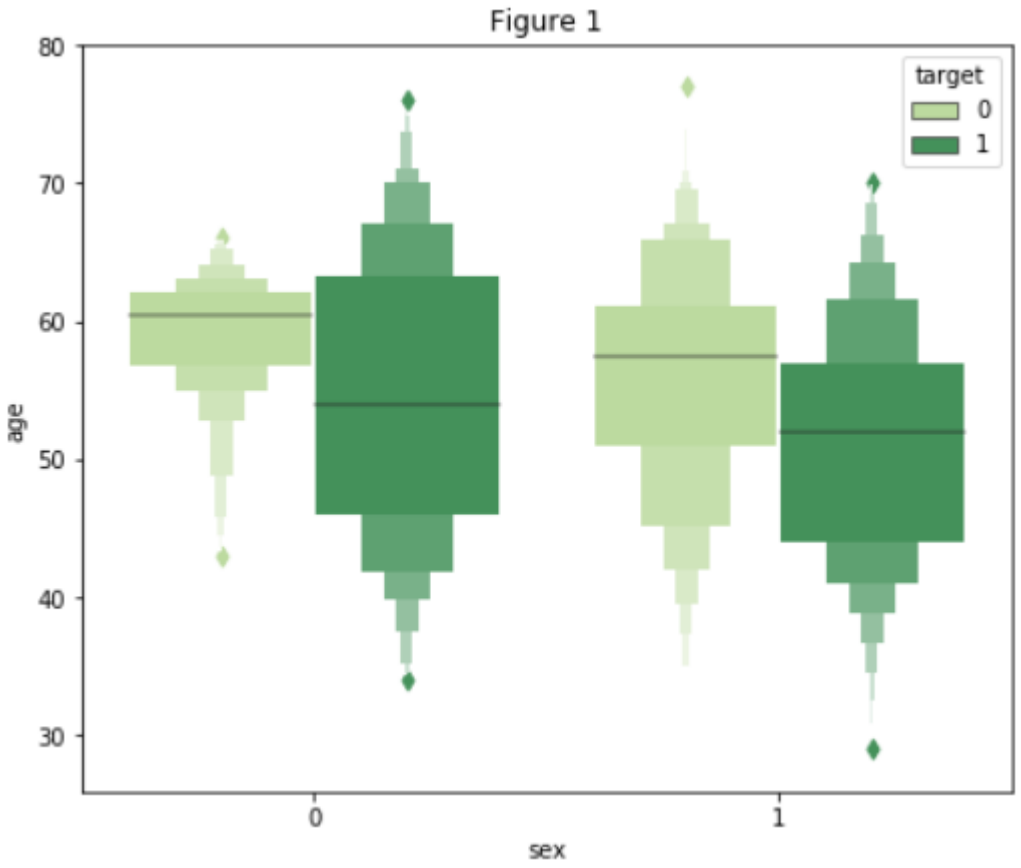
41-50 年龄段的人群患有心脏病的人数较多，而 60-70 岁的去检查的人中，患病人数相对还会较少。

推测一下，40-50 的中年人可能忙于工作等事情，去检查的人可能大多都是感觉身体出现了不适的情况，才去检查，所以患病人数占比较大。

02 性别与心脏病

```
fig,ax=plt.subplots(figsize=(16,6))
plt.subplot(121)
```

```
s1=sns.boxenplot(x='sex',y='age',hue='target',data=df,palette='YlGn')
s1.set_title("Figure 1")
```



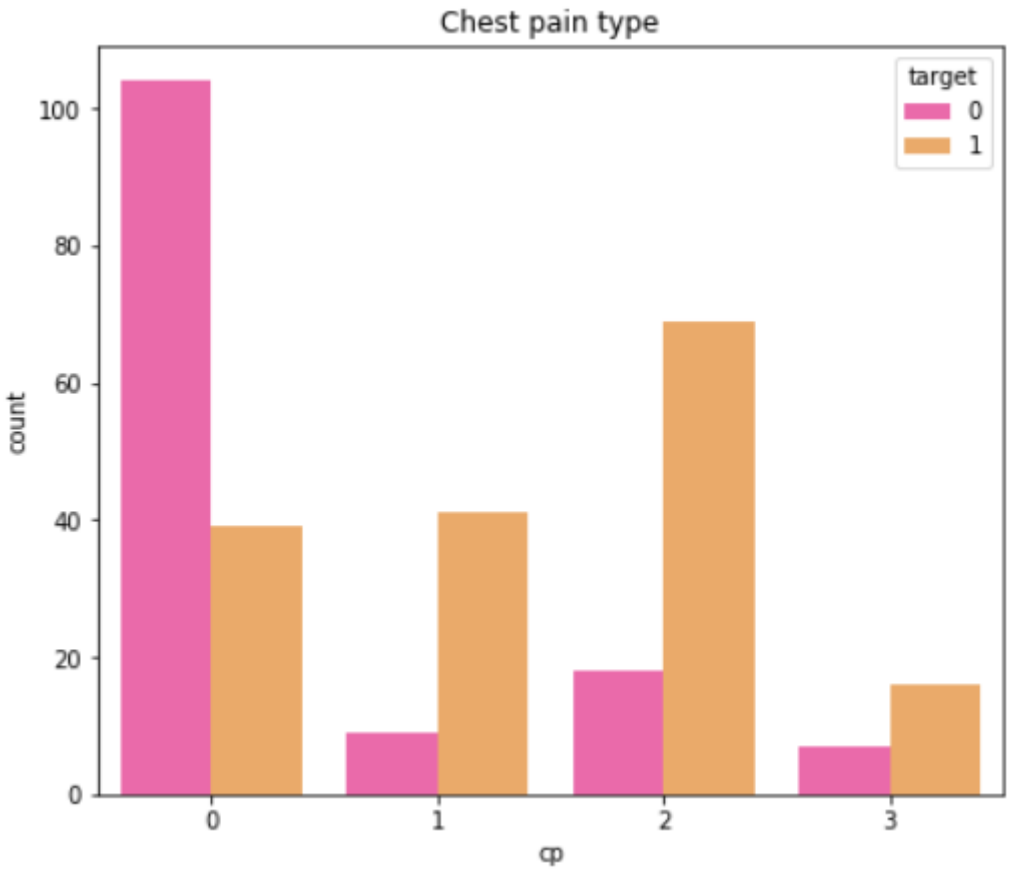
通过箱线图可以看到大多数女性患者年龄在 40-70 左右，男性患者大多在 40-60 岁，是否说明老年男性患病的可能性比老年女性患病的可能性小呢？

03 胸痛和心脏病

胸痛和心脏病在刚才的热图上看，相关性相对于其他因素来说是比较大的，那我们具体看一下。

```
fig,ax=plt.subplots(figsize=(24,6))
plt.subplot(131)
```

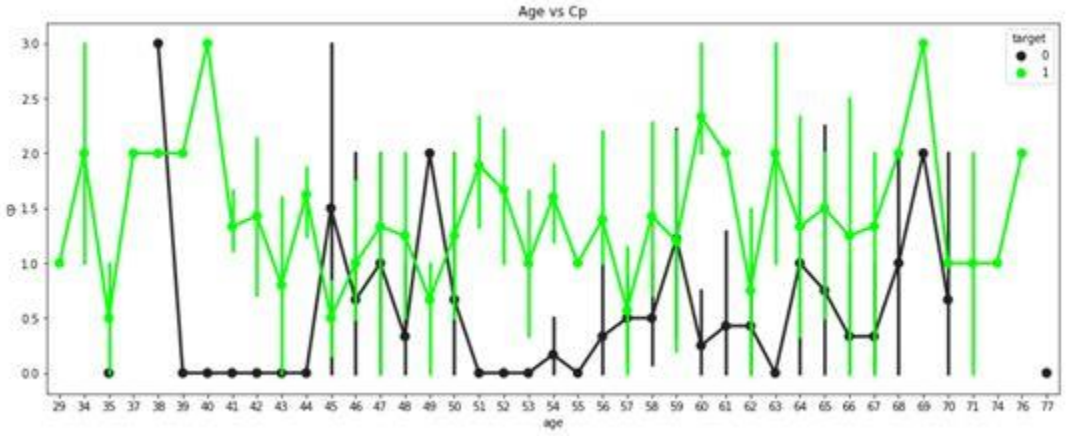
```
x1=sns.countplot(x='cp',data=df,hue='target',palette='spring')
x1.set_title('Chestpain type')
```



从图上可以看出，胸痛 2 型患有心脏病的概率最大。胸痛和不胸痛来看，胸痛的时候患有心脏病的概率大。

看一下年龄和胸痛有没有关系？

```
fig,ax=plt.subplots(figsize=(16,6))
sns.pointplot(x='age',y='cp',data=df,color='Lime',hue='target')
plt.title('Age vs Cp')
```

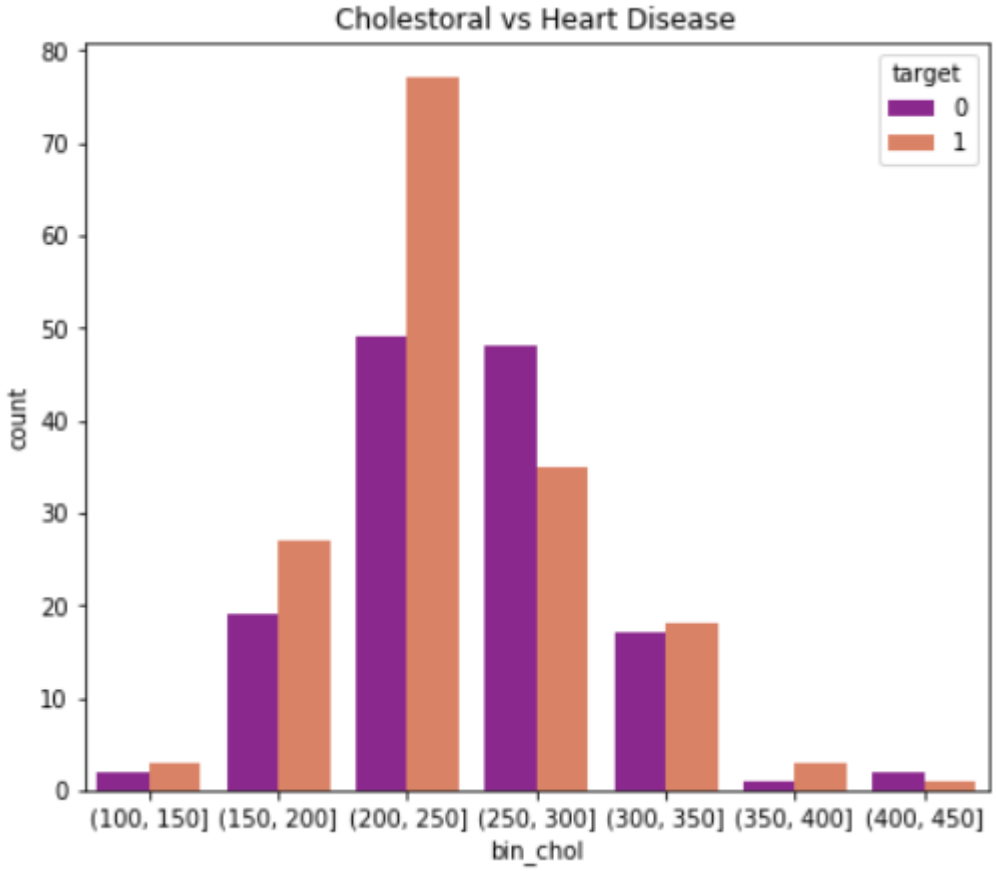


从这里看心脏病患者都具有较高的 cp 值，但是在 45-49 这段比较反常。

04 血清胆固醇的含量与心脏病

下面分析一下血清胆固醇的含量与心脏病的关联，从导致疾病的机理来看，长期的高血脂会造成血管内皮损伤，血管通透性增加等情况出现，增大了患有心血管疾病的可能性。

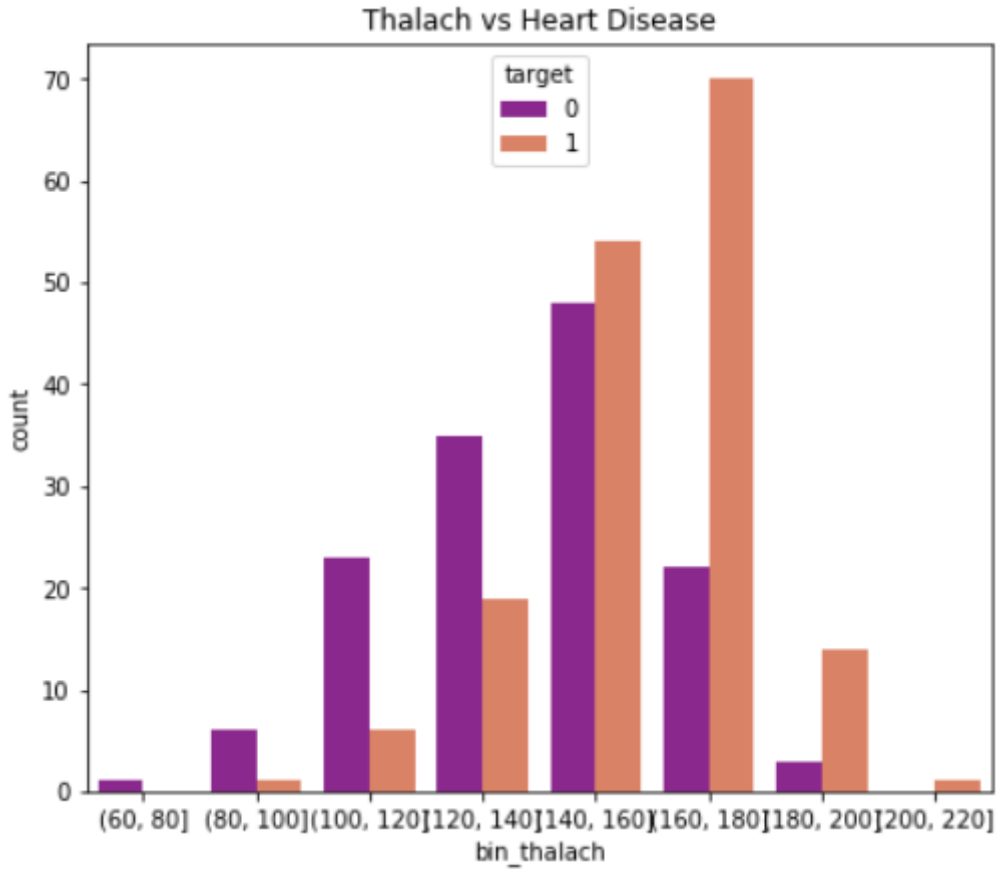
```
fig,ax=plt.subplots(figsize=(24,6))
plt.subplot(1,3,  2)
cho_bins= [100,150,200,250,300,350,400,450]
df['bin_chol']=pd.cut(df['chol'],bins=cho_bins)
g2=sns.countplot(x='bin_chol',data=df,hue='target',palette='plasma')
g2.set_title("Cholestoralvs Heart Disease")
```



血清胆固醇含量在 201-250 之间心脏病患病人数较多，而高于 250 后心脏病患病人数反而下降。

05 心率与心脏病

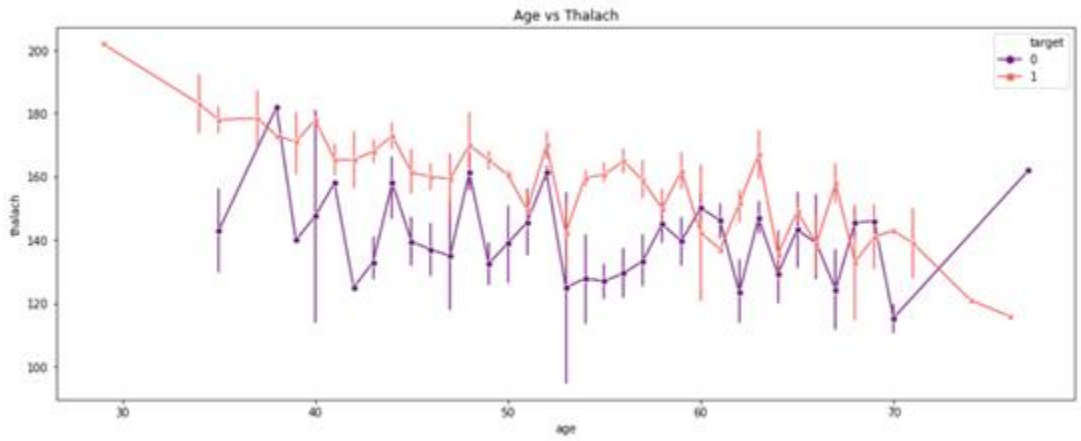
```
fig,ax=plt.subplots(figsize=(24,6))
plt.subplot(1,  3,  3)
thal_bins = [60,80,100,120,140,160,180,200,220]
df['bin_thalach']=pd.cut(df['thalach'], bins=thal_bins)
g3=sns.countplot(x='bin_thalach',data=df,hue='target',palette='plasma')
g3.set_title("Thalach vs Heart Disease")
```



从最大心率可以看出，最大心率在 140-180 之间的人，患有心脏病的风险最高。

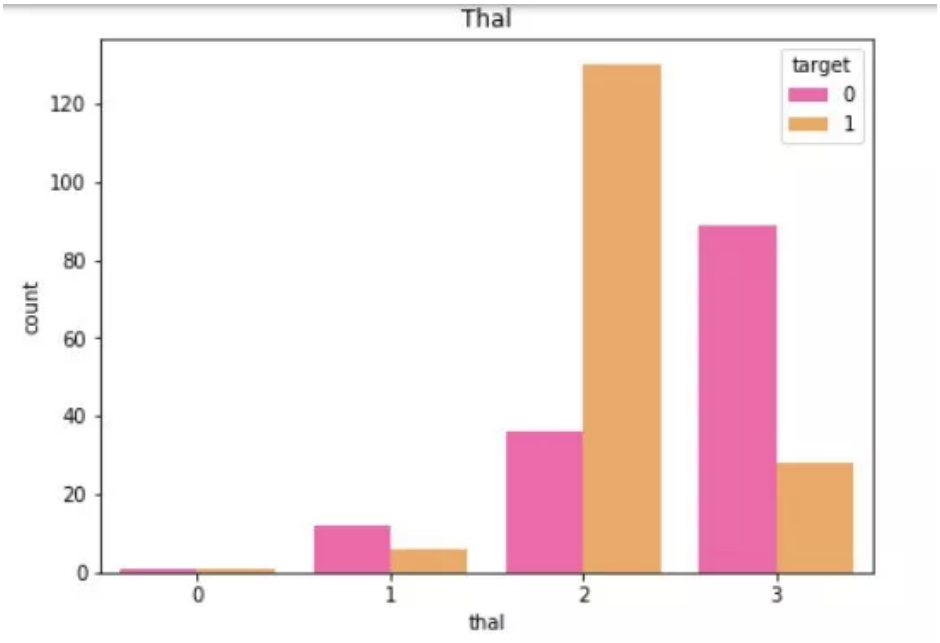
年龄和最大心率有没有啥相关性呢？

```
fig,ax=plt.subplots(figsize=(16,6))
sns.lineplot(y='thalach',x='age',data=df,hue="target",style='target',palette='
magma',markers=True,dashes=False,err_style="bars", ci=68)
plt.title('Agevs Thalach')
```



从这个图可以看出来，每个年龄段患有心脏病的患者，最大心率都相对较高。随着年龄的增大，最大心率会下降，这可能和年龄增加，身体机能下降有关。

06 thal 字段数据

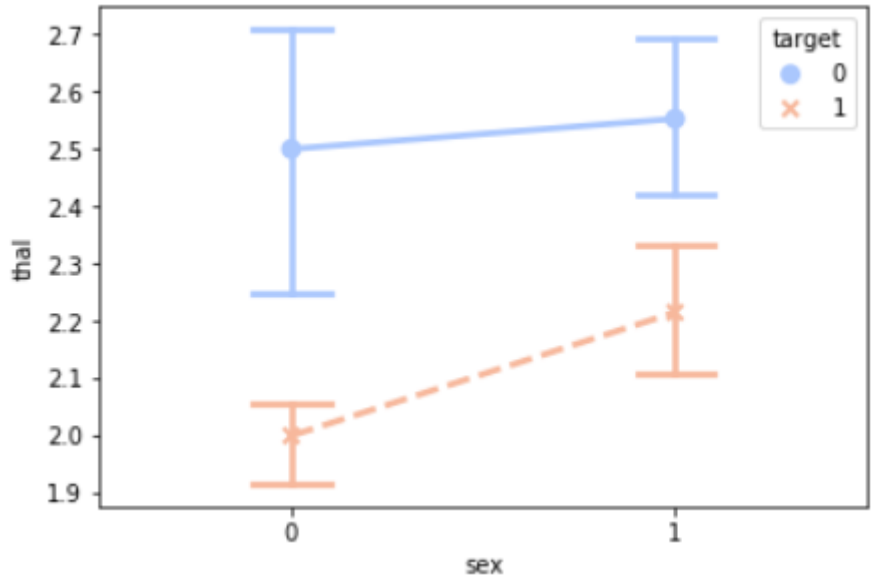


根据所给信息 thal 字段有 3 7 9 三个数值分别代表正常，固定缺陷，可逆缺陷，但是 thal 数值中并未出现这三个值，反之有 0,1,2,3 四种值。这让我感觉很疑惑。可能是数据源提供的信息出现些问题？

但是根据 thal 的图来看 thal 值为 2 的患有心脏病的概率最高。

我们看一下男性患者和女性患者和 thal 值有什么相关性没。

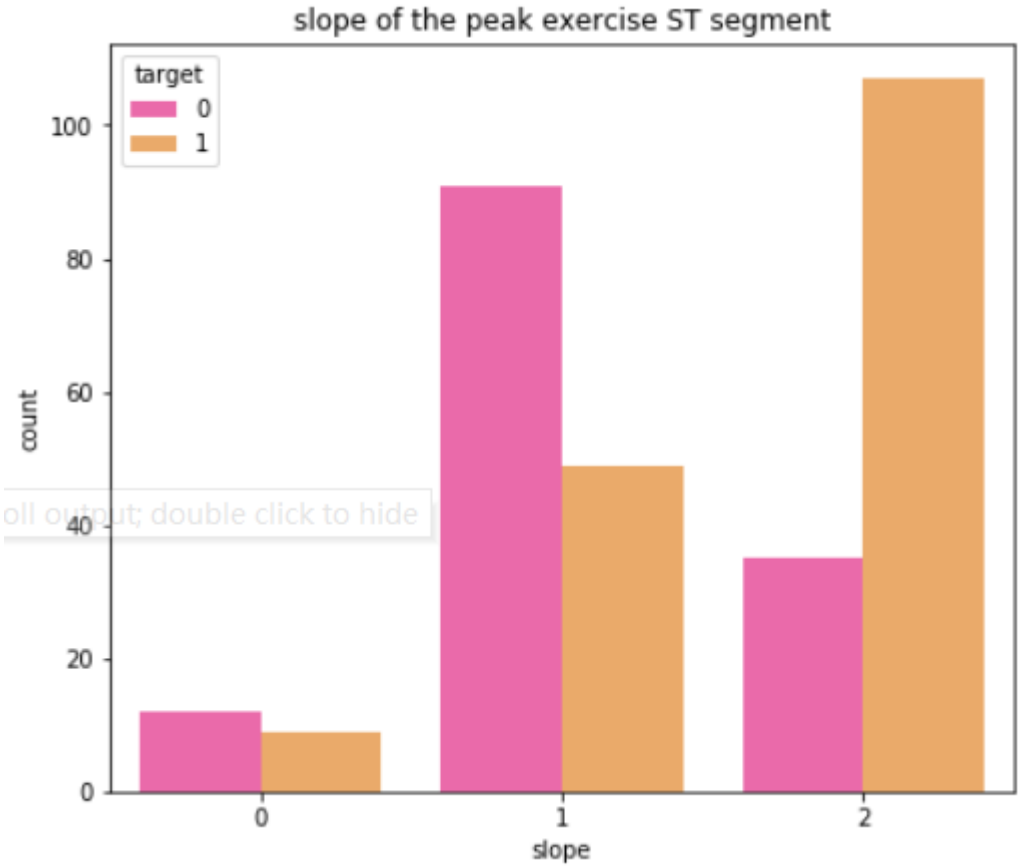
```
sns.pointplot(x='sex',y='thal',data=df,hue='target',markers=["o","x"],linestyles=["-","--"],capsize=.2,palette='coolwarm')
```



不管是男性还是女性，未患病的 thal 值均较高。而患病者中，男性的 thal 值会比女性的 thal 值高。

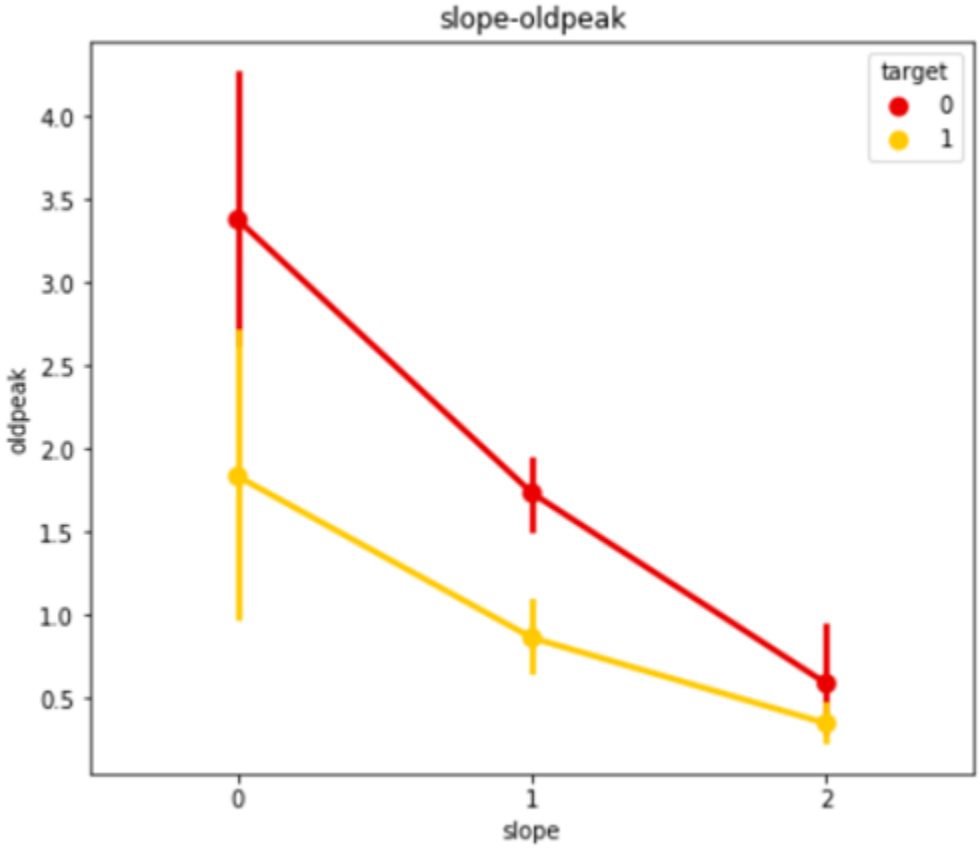
slope 值代表运动后 ST 段的斜率，心电图中，如果 ST 段异常变大说明心脏可能出现了心肌缺血的问题，先看下 slope 值和患病有什么关系。

```
fig,ax=plt.subplots(figsize=(24,6))
plt.subplot(133)
x3=sns.countplot(x='slope',data=df,hue='target',palette='spring')
x3.set_title('slopeof the peak exercise ST segment')
```



可以看到 slope 值为 2 的人群患心脏病的可能性比较大。

看下运动后 ST 段的斜率和相对于休息时候 ST 段抑制有没有什么相关性



我们可以看到随着 slope 值的减小，oldpeak 值也逐渐减小，并且我们可以发现心脏病患者的 oldpeak 值均相对健康人群 oldpeak 值低。

好了，我们分析了年龄、cp、thalach、slope 这四个字段以及它们和别的字段结合与患病的相关性。

本文参考了 kaggle 上大神的代码~和分析思路~云感谢大佬们~