注释：

红色标注的为值得注意和探讨的，bullet point 后面的红色表明是否有显著性，以及哪些指标有显著性。

* **Multicollinearity between numerical variables:**



It indicates that the pairs *(age, thalach), (oldpeak, thalach)* and *(age, ca)* have relatively higher linear relationship.

* **Distribution of Heart Disease:**



The dataset is relatively balanced.

* **Distribution of Heart Disease(age): (53-68)**



People who are older than 50, younger than 70 have higher heart disease risk. (分组，再探索)



It indicates that ages between 53 and 68 have higher significance with heart disease.

* 年龄再探索，两到三组 （成岳）
* **Distribution of Sex: (√)**



It indicates that numbers of females and males are not equal, so in the data exploration we should avoid comparing volume of males with females.

* **Distribution of Heart Disease(sex): (√)**



It indicates that the morbidity (患病率) of males is higher than that of females, so sex is an significant feature.

* **Distribution of Heart Disease(cp): (**Asymptomatic**)**



It indicates that Asymptomatic is a significant indicator of heart disease.

分为两类 asymptotic

* **Trestbps: (>143)**



Not very significant.



It indicates that trestbps larger than 143 is a significant indicator of Heart Disease.

分为143 和小于143

* **~~Chol: (x)~~**





They both show no significance.

* **~~Distribution of Heart Disease(fbs): (x)~~**



No significance.

* **Distribution of Heart Disease(restecg): (ST-T + Left)**



It might be better if combining the ST-T and Left together.

分成两组

* **Thalach: (√) \*处理成（测试值-理论值）**thalach with 220-age



It should have negative significance. But is has linear relationship with age, so we may be careful including both of the two indicators. Try to compare thalach with 220-age.

* **Distribution of Heart Disease(exang): (√)**

****

Have positive significance.

* **Oldpeak: (√) 留成数值 或者因子 为1  
  **

****

It has positive effect on heart disease.

* **Distribution of Heart Disease(slope): (non-upsloping)**



Slope 也分为两类，up 和非正常的

* **Distribution of Heart Disease(ca): (√)**



Positive significance.

* **Distribution of Heart Disease(thal): (Fixed+Reversable)**



Fixed+Reversable seem more significant.

Normal + abnormal 两类

补充：

查全率和查准率的问题

\*数据集选择

1. 原始数据 dataset1
2. 原始数据+缩放 dataset2
3. 全因子数据 dataset3
4. 原始数据 dataset1 – chol – fbs : dataset4
5. 原始数据+缩放 dataset2 – chol – fbs : dataset5
6. 全因子数据 dataset3 – chol – fbs : dataset6