数据分析与实践 实验三 实验报告

崔士强 PB22151743

部分题目没有输出，故留空.

任务一

Q1

Q2

```
         id diagnosis  radius_mean  texture_mean  perimeter_mean  area_mean  \
0    842302         M        17.99         10.38          122.80     1001.0
1    842517         M        20.57         17.77          132.90     1326.0
2  84300903         M        19.69         21.25          130.00     1203.0
3  84348301         M        11.42         20.38           77.58      386.1
4  84358402         M        20.29         14.34          135.10     1297.0
5    843786         M        12.45         15.70           82.57      477.1
6    844359         M        18.25         19.98          119.60     1040.0
7  84458202         M        13.71         20.83           90.20      577.9
8    844981         M        13.00         21.82           87.50      519.8
9  84501001         M        12.46         24.04           83.97      475.9

   smoothness_mean  compactness_mean  concavity_mean  concave points_mean  \
0          0.11840           0.27760         0.30010              0.14710
1          0.08474           0.07864         0.08690              0.07017
2          0.10960           0.15990         0.19740              0.12790
3          0.14250           0.28390         0.24140              0.10520
4          0.10030           0.13280         0.19800              0.10430
5          0.12780           0.17000         0.15780              0.08089
6          0.09463           0.10900         0.11270              0.07400
7          0.11890           0.16450         0.09366              0.05985
8          0.12730           0.19320         0.18590              0.09353
9          0.11860           0.23960         0.22730              0.08543

   ...  radius_worst  texture_worst  perimeter_worst  area_worst  \
0  ...         25.38          17.33           184.60      2019.0
1  ...         24.99          23.41           158.80      1956.0
2  ...         23.57          25.53           152.50      1709.0
3  ...         14.91          26.50            98.87       567.7
4  ...         22.54          16.67           152.20      1575.0
5  ...         15.47          23.75           103.40       741.6
6  ...         22.88          27.66           153.20      1606.0
7  ...         17.06          28.14           110.60       897.0
8  ...         15.49          30.73           106.20       739.3
9  ...         15.09          40.68            97.65       711.4

   smoothness_worst  compactness_worst  concavity_worst  concave points_worst  \
0            0.1622             0.6656           0.7119                0.2654
1            0.1238             0.1866           0.2416                0.1860
2            0.1444             0.4245           0.4504                0.2430
3            0.2098             0.8663           0.6869                0.2575
4            0.1374             0.2050           0.4000                0.1625
5            0.1791             0.5249           0.5355                0.1741
6            0.1442             0.2576           0.3784                0.1932
7            0.1654             0.3682           0.2678                0.1556
8            0.1703             0.5401           0.5390                0.2060
9            0.1853             1.0580           1.1050                0.2210

   symmetry_worst  fractal_dimension_worst
0          0.4601                  0.11890
1          0.2750                  0.08902
2          0.3613                  0.08758
3          0.6638                  0.17300
4          0.2364                  0.07678
5          0.3985                  0.12440
6          0.3063                  0.08368
7          0.3196                  0.11510
8          0.4378                  0.10720
9          0.4366                  0.20750

[10 rows x 32 columns]
```

Q3

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 32 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   id                       569 non-null    int64  
 1   diagnosis                569 non-null    object 
 2   radius_mean              569 non-null    float64
 3   texture_mean             569 non-null    float64
 4   perimeter_mean           569 non-null    float64
 5   area_mean                569 non-null    float64
 6   smoothness_mean          568 non-null    float64
 7   compactness_mean         569 non-null    float64
 8   concavity_mean           569 non-null    float64
 9   concave points_mean      569 non-null    float64
 10  symmetry_mean            569 non-null    float64
 11  fractal_dimension_mean   567 non-null    float64
 12  radius_se                569 non-null    float64
 13  texture_se               567 non-null    float64
 14  perimeter_se             569 non-null    float64
 15  area_se                  569 non-null    float64
 16  smoothness_se            569 non-null    float64
 17  compactness_se           568 non-null    float64
 18  concavity_se             568 non-null    float64
 19  concave points_se        569 non-null    float64
 20  symmetry_se              569 non-null    float64
 21  fractal_dimension_se     568 non-null    float64
 22  radius_worst             568 non-null    float64
 23  texture_worst            569 non-null    float64
 24  perimeter_worst          569 non-null    float64
 25  area_worst               569 non-null    float64
 26  smoothness_worst         568 non-null    float64
 27  compactness_worst        569 non-null    float64
 28  concavity_worst          569 non-null    float64
 29  concave points_worst     569 non-null    float64
 30  symmetry_worst           569 non-null    float64
 31  fractal_dimension_worst  569 non-null    float64
dtypes: float64(30), int64(1), object(1)
memory usage: 142.4+ KB
```

Q4

Q5

Q6

Q7

```
diagnosis
B    355
M    205
Name: count, dtype: int64
```

Q8

Q9

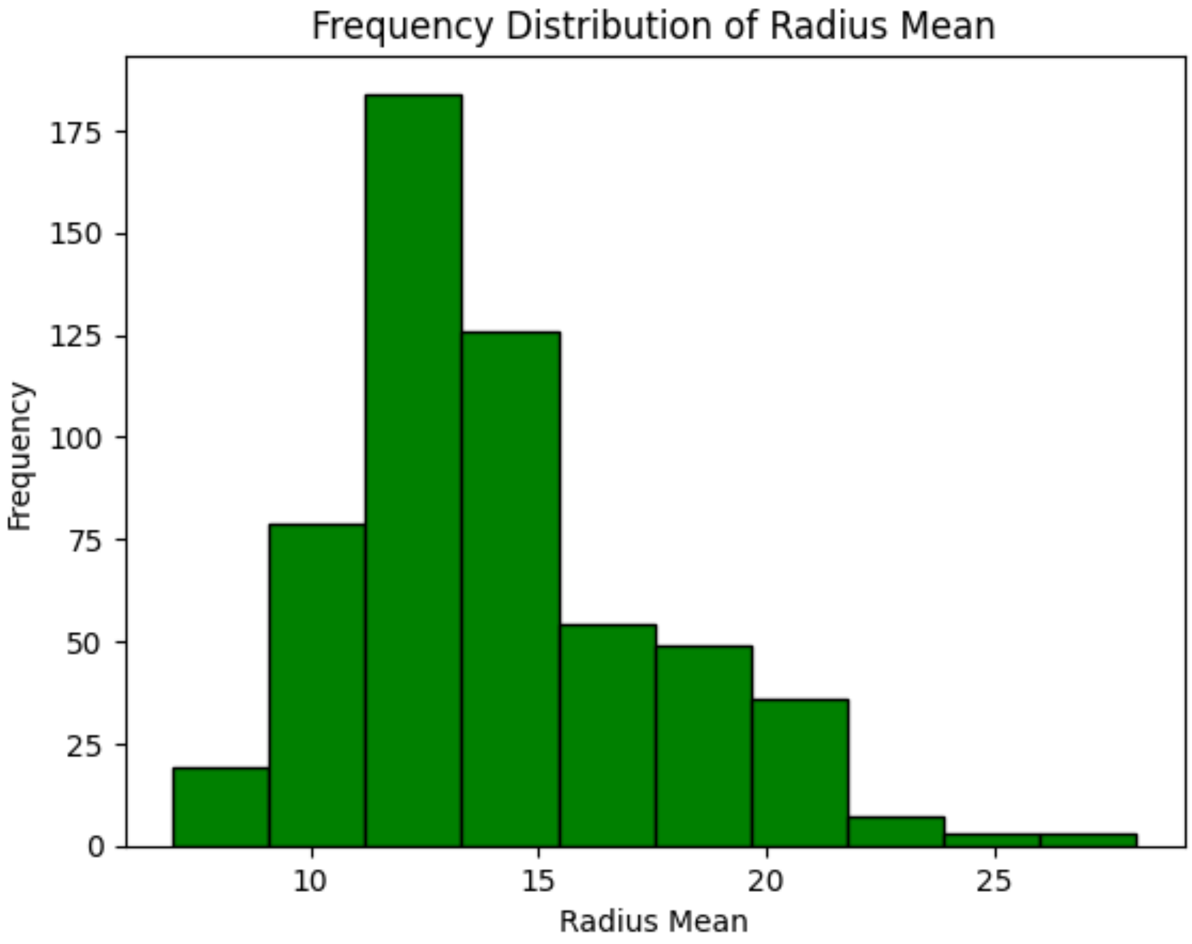
```
       radius_mean  texture_mean  perimeter_mean    area_mean  smoothness_mean
count   560.000000    560.000000      560.000000   560.000000       560.000000
mean     14.074302     19.271750       91.595857   649.643929         0.096281
std       3.491064      4.319015       24.048329   347.451287         0.014088
min       6.981000      9.710000       43.790000   143.500000         0.052630
25%      11.677500     16.157500       74.967500   418.325000         0.086290
50%      13.275000     18.825000       85.980000   544.050000         0.095785
75%      15.750000     21.802500      103.725000   775.775000         0.105100
max      28.110000     39.280000      188.500000  2501.000000         0.163400
```

Q10

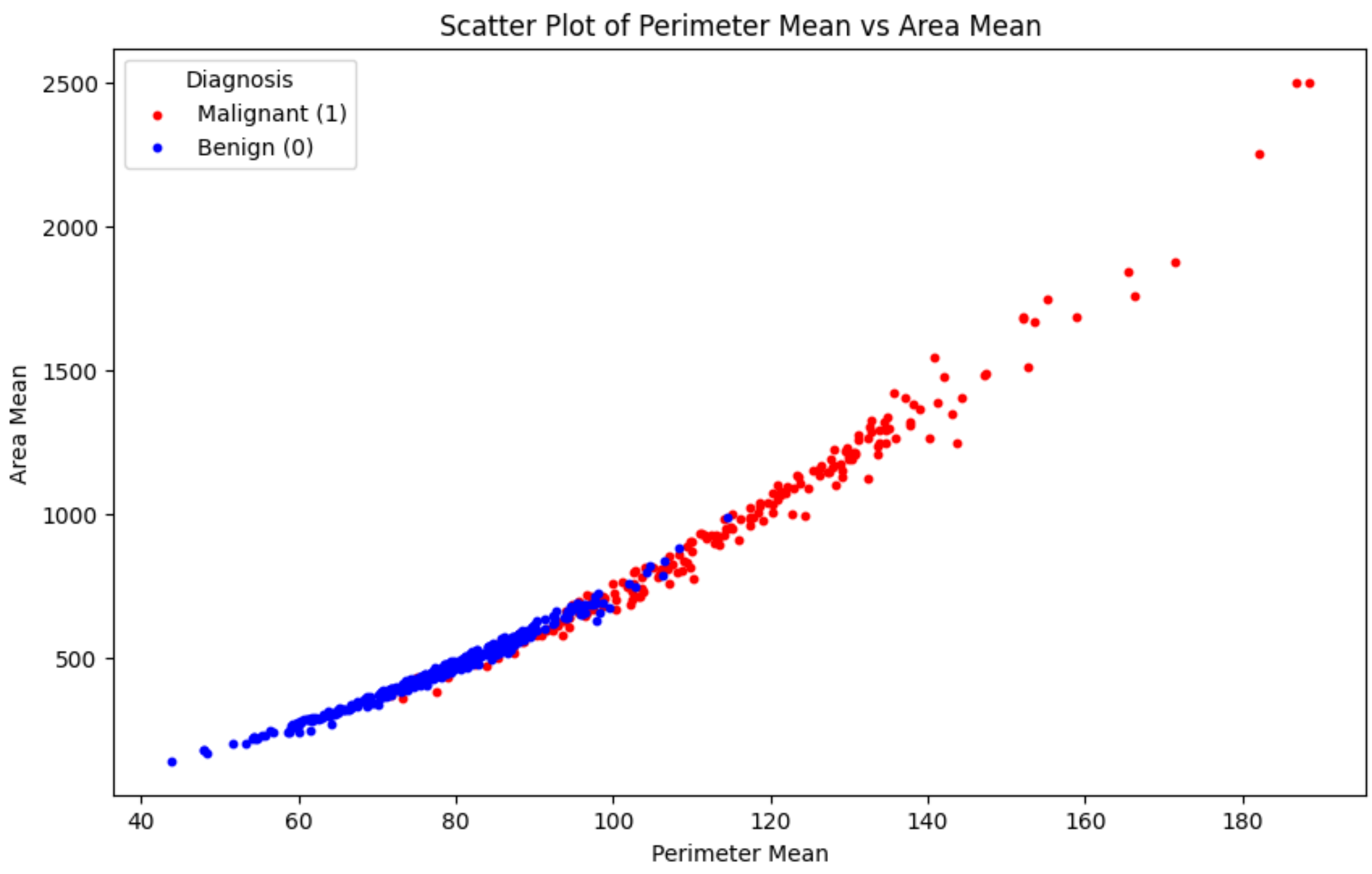
```
           radius_mean  texture_mean  perimeter_mean  area_mean  \
diagnosis
0             0.146810      0.223512        0.151502   0.290776
1             0.182084      0.175314        0.187713   0.373423

           smoothness_mean  compactness_mean  concavity_mean  \
diagnosis
0                 0.145493          0.421460        0.943919
1                 0.124034          0.370451        0.467383

           concave points_mean  symmetry_mean  fractal_dimension_mean  ...  \
diagnosis                                                              ...
0                     0.619809       0.142692                0.107291  ...
1                     0.387497       0.143382                0.121587  ...

           radius_worst  texture_worst  perimeter_worst  area_worst  \
diagnosis
0              0.148098       0.234088         0.155623    0.292836
1              0.200238       0.186279         0.205549    0.415632

           smoothness_worst  compactness_worst  concavity_worst  \
diagnosis
0                  0.160242           0.505433         0.846008
1                  0.153082           0.455924         0.407353

           concave points_worst  symmetry_worst  fractal_dimension_worst
diagnosis
0                      0.481809        0.154683                 0.173924
1                      0.250869        0.232936                 0.236436

[2 rows x 30 columns]
```

任务二

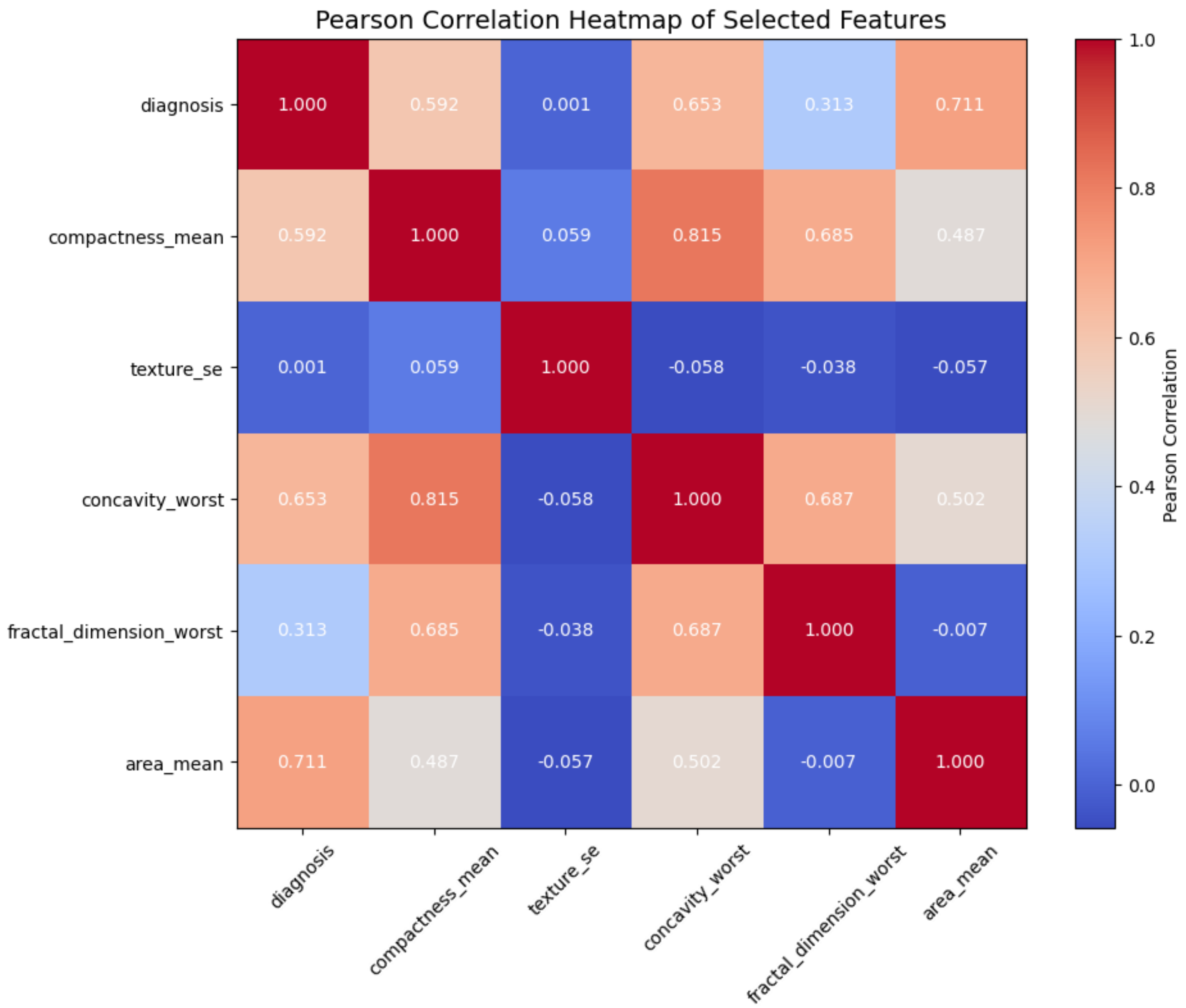
Q1



Q2


Scatter Plot of Perimeter Mean vs Area Mean
Diagnosis
● Malignant (1)
● Benign (0)
Area Mean
2500
2000
1500
1000
500
Perimeter Mean
40
60
80
100
120
140
160
180

Q3


Pearson Correlation Heatmap of Selected Features
diagnosis
compactness_mean
texture_se
concavity_worst
fractal_dimension_worst
area_mean
1.000
0.592
0.001
0.653
0.313
0.711
0.592
1.000
0.059
0.815
0.685
0.487
0.001
0.059
1.000
-0.058
-0.038
-0.057
0.653
0.815
-0.058
1.000
0.687
0.502
0.313
0.685
-0.038
0.687
1.000
-0.007
0.711
0.487
-0.057
0.502
-0.007
1.000
diagnosis
compactness_mean
texture_se
concavity_worst
fractal_dimension_worst
area_mean
Pearson Correlation
1.0
0.8
0.6
0.4
0.2
0.0

任务三

Q1

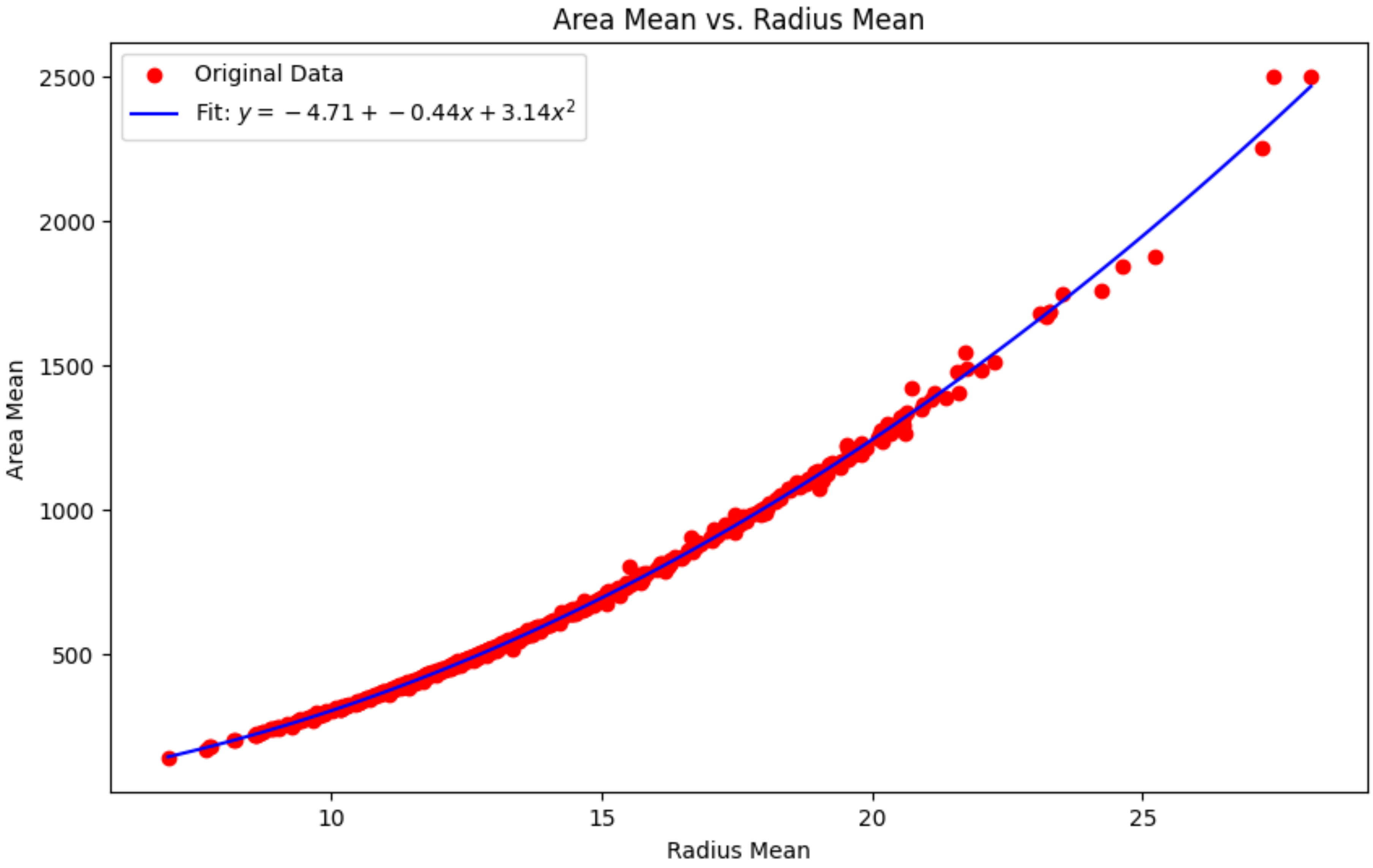
```
[-4.70867951 -0.44260792  3.14186228]
```

Q2

```
[ 3.14186228 -0.44260792 -4.70867951]
```

可以发现得到的结果与Q1中相等.

Q3



- 从特征含义来看，两个变量分别是圆的面积和半径，理论上应当是二次关系.
- 从散点分布来看，散点呈现曲线分布.

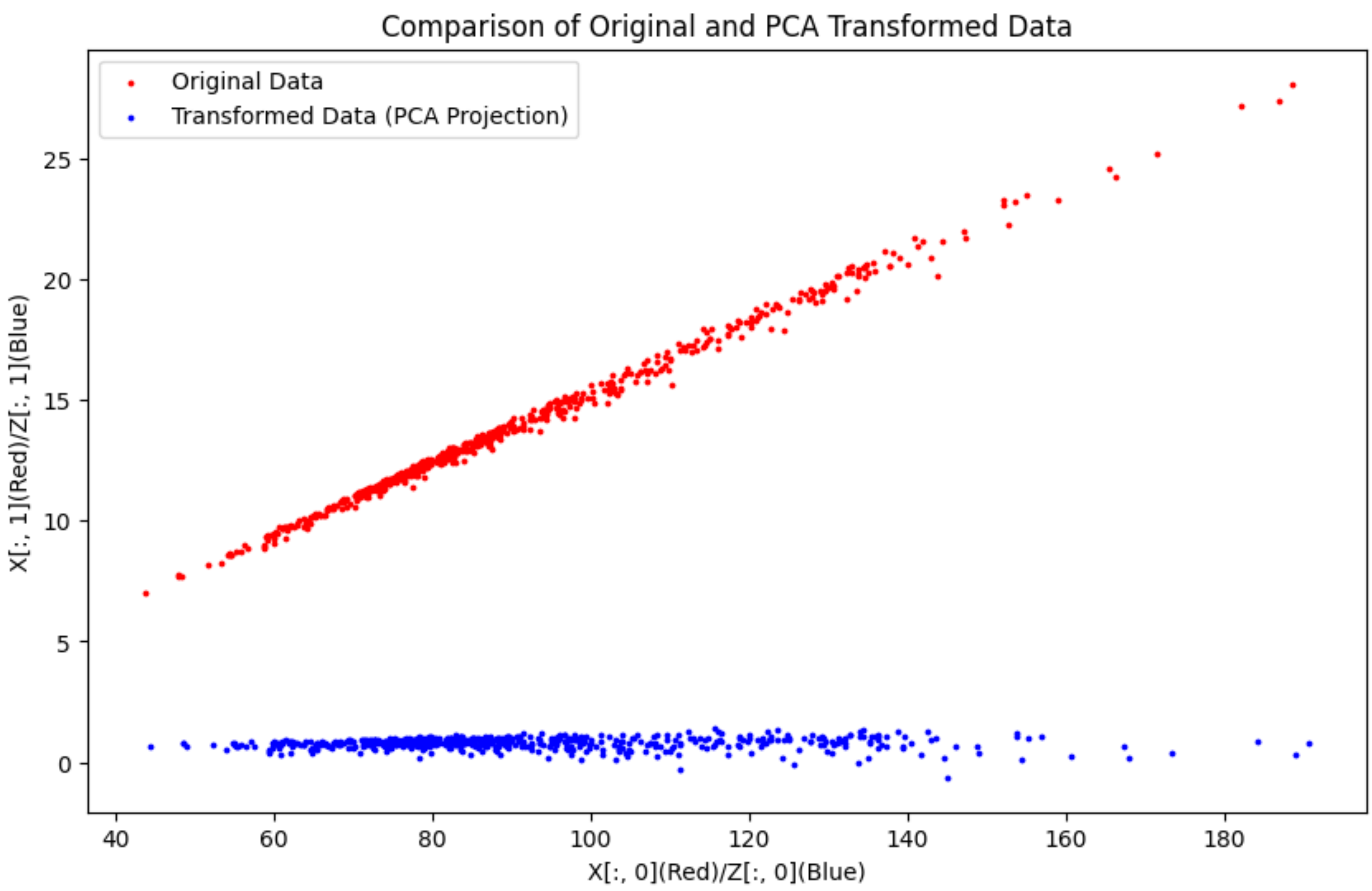
综合考虑这两个方面，线性拟合并不适用.

任务四

Q1

```
corX: [[578.32210982  83.77072122]
 [ 83.77072122  12.18753099]]
eigV: [5.90457503e+02 5.21373729e-02]
eigMat: [[ 0.98966947 -0.14336785]
 [ 0.14336785  0.98966947]]
Orthogonality: True
```

Q2



从图中可以看到，原数据呈现出两个维度，经过处理后被投射到一维.

Q3

```
[[ 5.90457503e+02 -4.52192448e-14]
 [-4.52192448e-14  5.21373729e-02]]
```

通过观察可以得到，这里得到的矩阵基本可以视作对角阵，对角元素即为 `eigV` .

任务五

Q1

由于 `diagnosis` 值将样本分为两个独立的组（恶性和良性），我们应使用成组检验，因为两组间没有重叠的个体.

Q2

```
Malignant Mean: 0.44671356097560977
Benign Mean: 0.1663615971830986
```

原假设 (H0): 恶性肿瘤（M）的 `concavity_worst` 平均值不大于良性肿瘤（B）的平均值.

备择假设 (H1): 恶性肿瘤（M）的 `concavity_worst` 平均值大于良性肿瘤（B）的平均值.

Q3

```
T-Statistic: 20.346631967479436
P-Value: 1.4640846763013996e-69
```

Q4

- T统计量(20.3466)显示了恶性肿瘤组的 `concavity_worst` 平均值与良性肿瘤组之间存在非常大的差异。
- P值(1.464×10^{-69})远远小于任何常用的显著性水平（如0.05、0.01甚至0.001）。这意味着观察到的差异极不可能是随机结果。这个极低的P值使我们有充足的理由拒绝原假设。

结论：恶性肿瘤的 `concavity_worst` 平均值显著大于良性肿瘤的平均值。