به نام خدا

سوال۲ فصل ۳)

داده‌های در نظر گرفته شده در این تمرین داده‌های pbc است که توسط آزمایشات کلینیک Mayo روی ریه ۴۱۸ بیمار مبتلا به PBC در سال‌های ۱۹۷۴ و ۱۹۸۴ انجام شده است که در طی آن بازه ده ساله به آن کلینیک مراجعه کرده اند. خلاصه ای از داده‌ها و متغیرهای مورد بحث در زیر نمایش داده شده است:

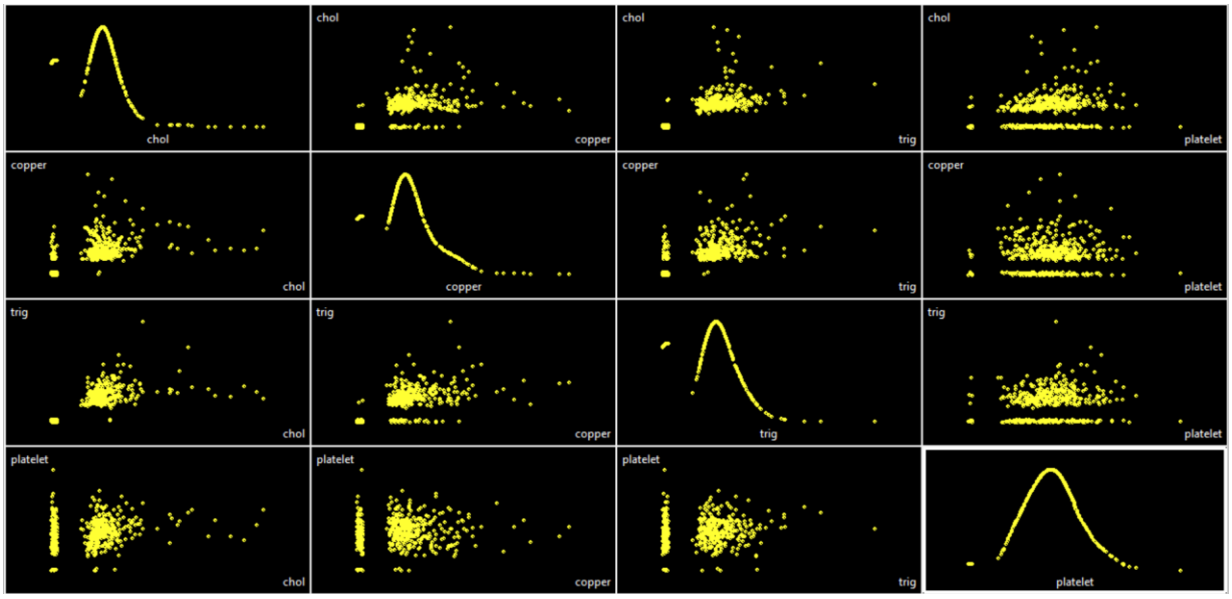
```
> summary(pbc)
      days           status         treatment          age              sex        
 Min.   :  41   Min.   :0.0000   Min.   :1.000   Min.   : 9598   Min.   :0.0000  
 1st Qu.:1093   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:15644   1st Qu.:1.0000  
 Median :1730   Median :0.0000   Median :1.000   Median :18628   Median :1.0000  
 Mean   :1918   Mean   :0.3852   Mean   :1.494   Mean   :18533   Mean   :0.8947  
 3rd Qu.:2614   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:21273   3rd Qu.:1.0000  
 Max.   :4795   Max.   :1.0000   Max.   :2.000   Max.   :28650   Max.   :1.0000  
                                 NA's   :106                                     
    ascites          hepatom          spiders           edema             bili        
 Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   : 0.300  
 1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.800  
 Median :0.00000   Median :1.0000   Median :0.0000   Median :0.0000   Median : 1.400  
 Mean   :0.07692   Mean   :0.5128   Mean   :0.2885   Mean   :0.1005   Mean   : 3.221  
 3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.: 3.400  
 Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :28.000  
 NA's   :106       NA's   :106      NA's   :106                                       
      chol           albumin          copper            alk              sgot        
 Min.   : 120.0   Min.   :1.960   Min.   :  4.00   Min.   :  289.0   Min.   : 26.35  
 1st Qu.: 249.5   1st Qu.:3.243   1st Qu.: 41.25   1st Qu.:  871.5   1st Qu.: 80.60  
 Median : 309.5   Median :3.530   Median : 73.00   Median : 1259.0   Median :114.70  
 Mean   : 369.5   Mean   :3.497   Mean   : 97.65   Mean   : 1982.7   Mean   :122.56  
 3rd Qu.: 400.0   3rd Qu.:3.770   3rd Qu.:123.00   3rd Qu.: 1980.0   3rd Qu.:151.90  
 Max.   :1775.0   Max.   :4.640   Max.   :588.00   Max.   :13862.4   Max.   :457.25  
 NA's   :134                      NA's   :108      NA's   :106       NA's   :106     
      trig           platelet      prothrombin        stage      
 Min.   : 33.00   Min.   : 62.0   Min.   : 9.00   Min.   :1.000  
 1st Qu.: 84.25   1st Qu.:188.5   1st Qu.:10.00   1st Qu.:2.000  
 Median :108.00   Median :251.0   Median :10.60   Median :3.000  
 Mean   :124.70   Mean   :257.0   Mean   :10.73   Mean   :3.024  
 3rd Qu.:151.00   3rd Qu.:318.0   3rd Qu.:11.10   3rd Qu.:4.000  
 Max.   :598.00   Max.   :721.0   Max.   :18.00   Max.   :4.000  
 NA's   :136      NA's   :11      NA's   :2       NA's   :6      
```

```
> attributes(pbc)
$names
 [1] "days"       "status"     "treatment"  "age"        "sex"        "ascites"    "hepatom"    "spiders"    "edema"      "bili"       "chol"      
[12] "albumin"    "copper"     "alk"        "sgot"       "trig"       "platelet"   "prothrombin" "stage"

$class
[1] "data.frame"
```

همانطور که مشاهده می‌شود مقادیر گمشده و دور افتاده در داده‌ها وجود دارد که در ادامه با به تصویر کشیدن داده‌ها می‌توان بهتر آن‌ها را مشاهده کرد.

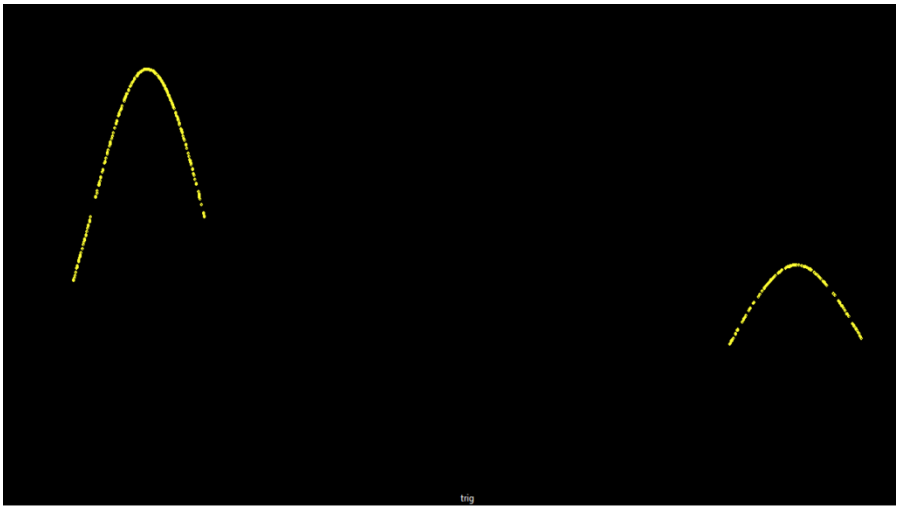
نمودار ماتریس پراکنش ۴ متغیر chol (مقدار کلسترول خون) ,copper (مقدار مس موجود در خون) , trig (مقدار تری گلیسیرید خون) و platelet (تعداد پلاکت خون) به صورت زیر می‌باشد.



با توجه به نمودار بالا ارتباط خطی بین این متغیرها وجود ندارد، همچنین در توزیع چگالی آن‌ها چولگی به سمت راست (بیشتر در سه متغیر اول) مشاهده می‌شود و به خاطره وجود داده‌های دور افتاده و داده‌های گمشده می‌باشد.

در ادامه، داده‌های گمشده را با مقدار مینیمم جانهی کرده و نمودارهای چگالی ۴ متغیر به صورت زیر بدست می‌آید.


copper


trig


platelet

همانطور که مشاهده می‌شود نمودارها به دو قسمت، که هر کدام نمایانگر یک کله قند می‌باشند تبدیل شده و از روی این می‌توان گفت داده‌های گمشده در نرمال نشدن توزیع نقش بسزایی دارند.

تعداد داده‌های گمشده برای متغیرهای پیوسته به صورت زیر است

```
    chol  albumin   copper alk.phos      ast     trig   protime
     134        0      108      106      106      136         2
platelet
      11
```

و با توجه به جدول زیر تعداد داده‌های گمشده در هر سطر و ستون مشاهده می‌شود برای مثال ۲۷۶ داده همگی مشاهده شده (recorded) هستند، ۲ داده در متغیر سوم در نظر گرفته شده در اینجا (copper) گمشده و ۲۸ داده در متغیرهای اول و ششم گمشده می‌باشند و ... . (مقادیر گمشده مشترک در بین چند متغیر را در سطرها می‌توان مشاهده کرد)

از طرفی می‌توان به این صورت به جدول نگاه کرد، در متغیر اول ۲۸+۹۹+۷ تعداد داده‌های گمشده می‌باشد که برابر با همان ۱۳۴ است که ۲۸ داده در بین دو متغیر اول و ششم، ۹۹ داده در بین متغیرهای اول، سوم تا ششم قرار دارد و ۷ داده دیگر نیز در بین همه متغیرها به جز متغیر دوم وجود دارد.

همچنین در متغیر دوم مقدار گمشده ای وجود ندارد و در متغیر سوم نیز به همین منوال، ۱۰۸ داده گمشده موجود می‌باشد.

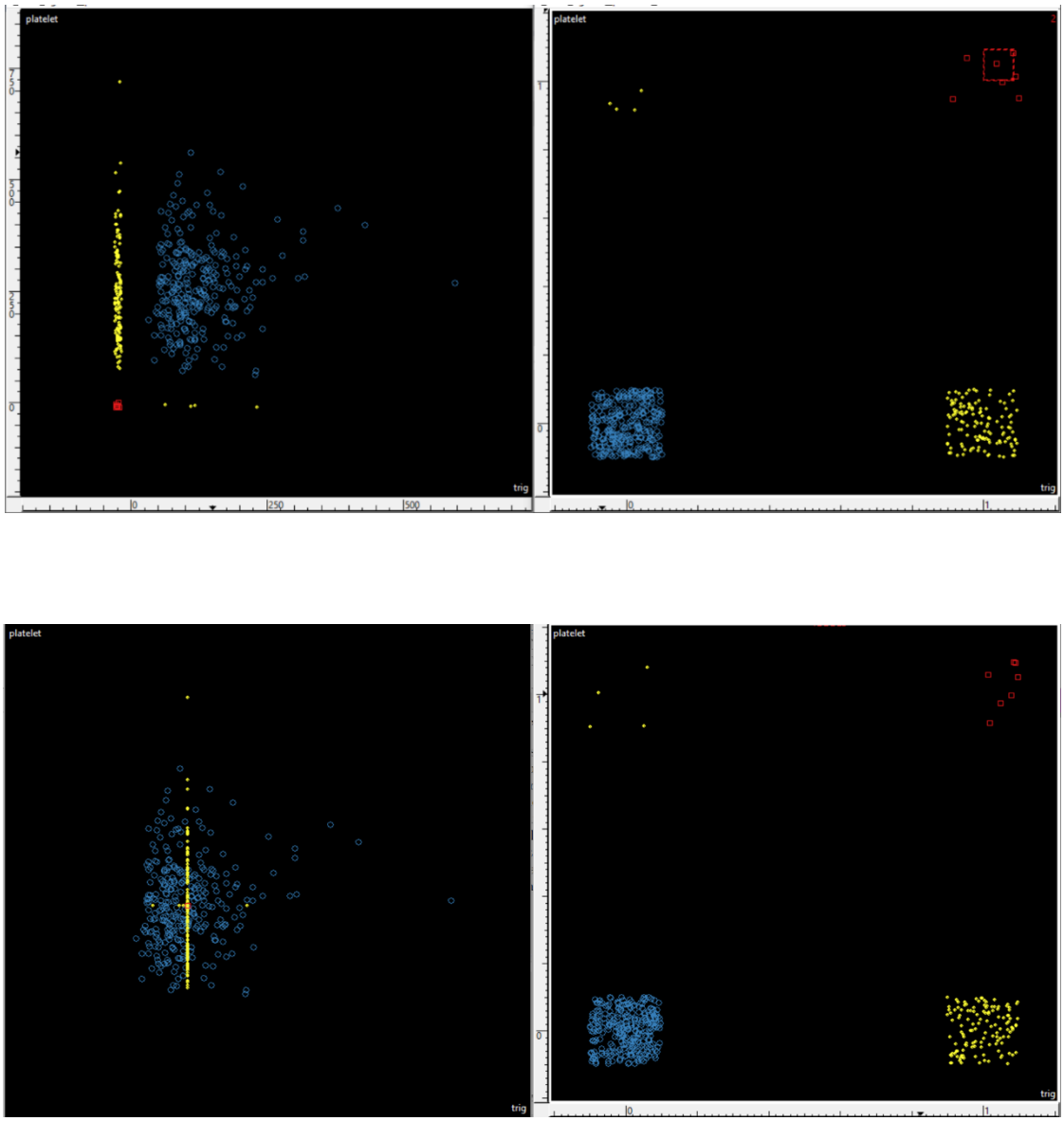
```
    [,1] [,2] [,3] [,4] [,5] [,6] [,7]
276    1    1    1    1    1    1    1
  2    1    1    0    1    1    1    1
  2    1    1    1    1    1    0    1
 28    0    1    1    1    1    0    1
 99    0    1    0    0    0    0    1
  4    1    1    1    1    1    1    0
  7    0    1    0    0    0    0    0
```

برای متغیر protime نیز به صورت جداگانه این کار انجام شده و همانطور که در زیر مشاهده می‌شود ۲ مقدار گمشده در آن وجود دارد.

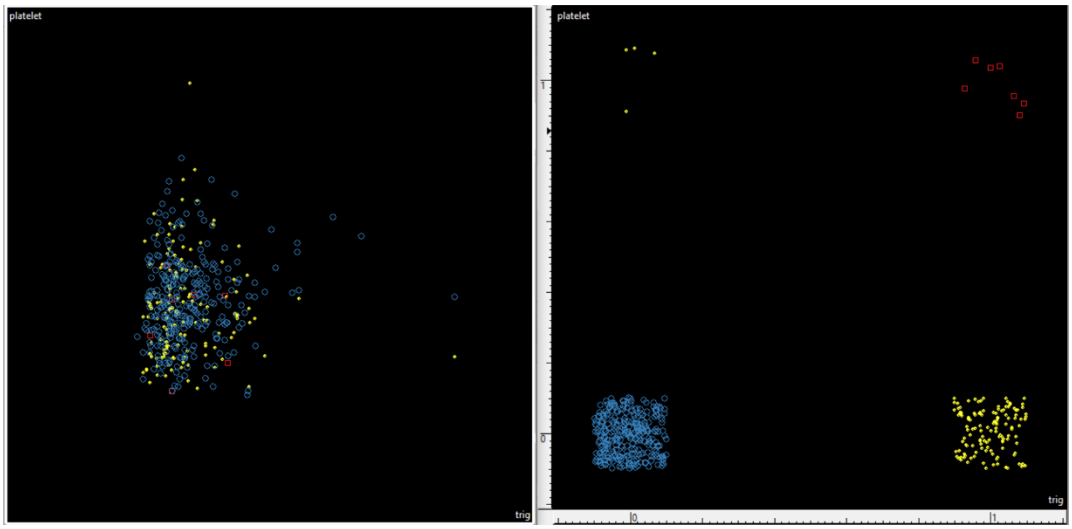
```
    [,1]
416    1
  2    0
```

(صفر در اینجا نمایانگر داده گمشده و یک نمایانگر داده مشاهده شده، می‌باشد)

ابتدا مقادیر گمشده را با مقدار مینیمم، میانگین و مقادیر تصادفی جانهی کرده و به همین ترتیب، نمودارهای پراکنش متغیر trig و platelet و missing plot به صورت زیر می‌باشند که در آن رنگ آبی نشان دهنده داده‌های مشاهده شده و رنگ زرد و قرمز نشان دهنده داده‌های گمشده هستند که بخش قرمز متعلق به داده‌های گمشده هر دو متغیر و داده‌های زرد متعلق به داده‌های گمشده یکی از اینها می‌باشد.

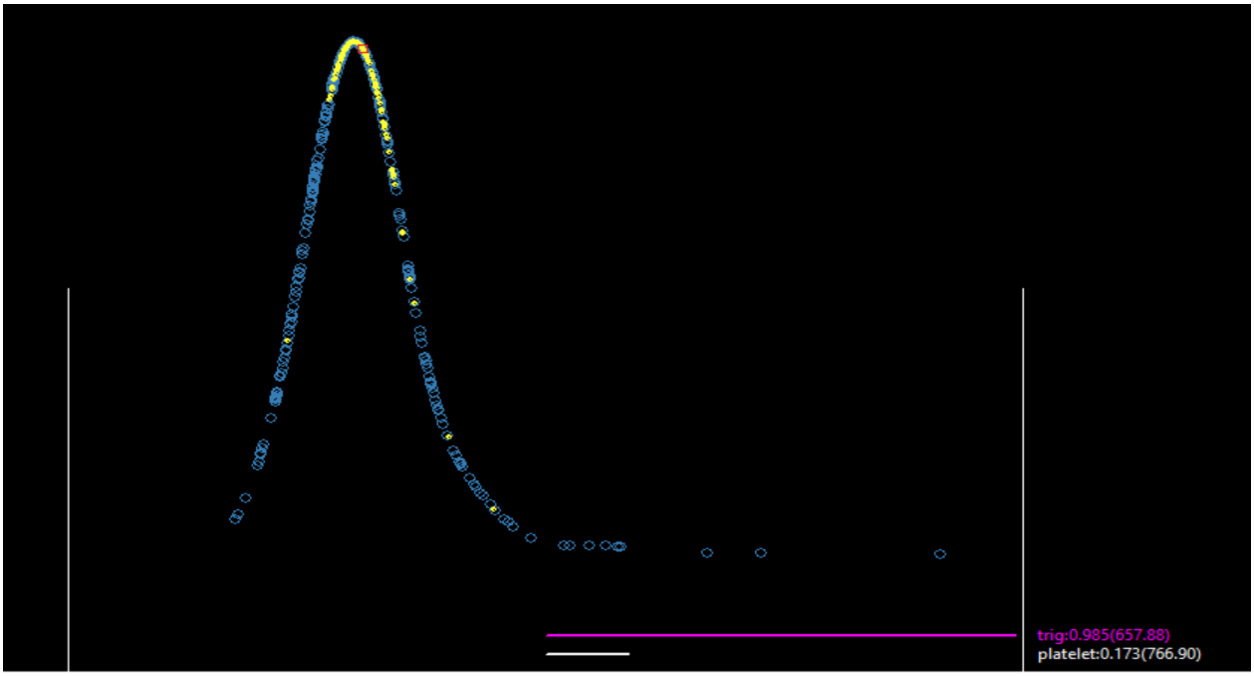


همانطور که مشاهده می‌شود در دو شکل اول چون جانهی با یک مقدار خاص (مینیم و میانگین) انجام شده، مقادیر بدست آمده روی یک خط قرار گرفته اند از همین رو در گام بعد با مقادیر تصادفی جانهی را انجام داده و نمودار زیر بدست آمده است.

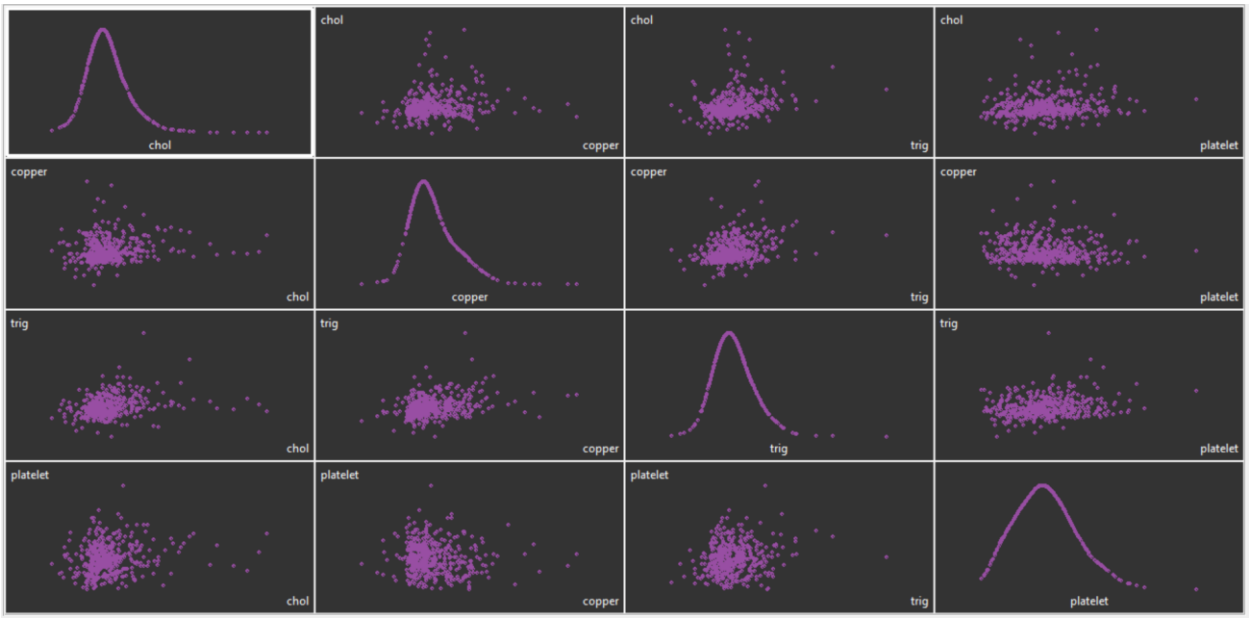


که همانطور که مشاهده می‌شود جانهی بهتر از دو حالت قبل انجام شده است و همچنان رابطه بین این دو متغیر خطی نمی‌باشد.

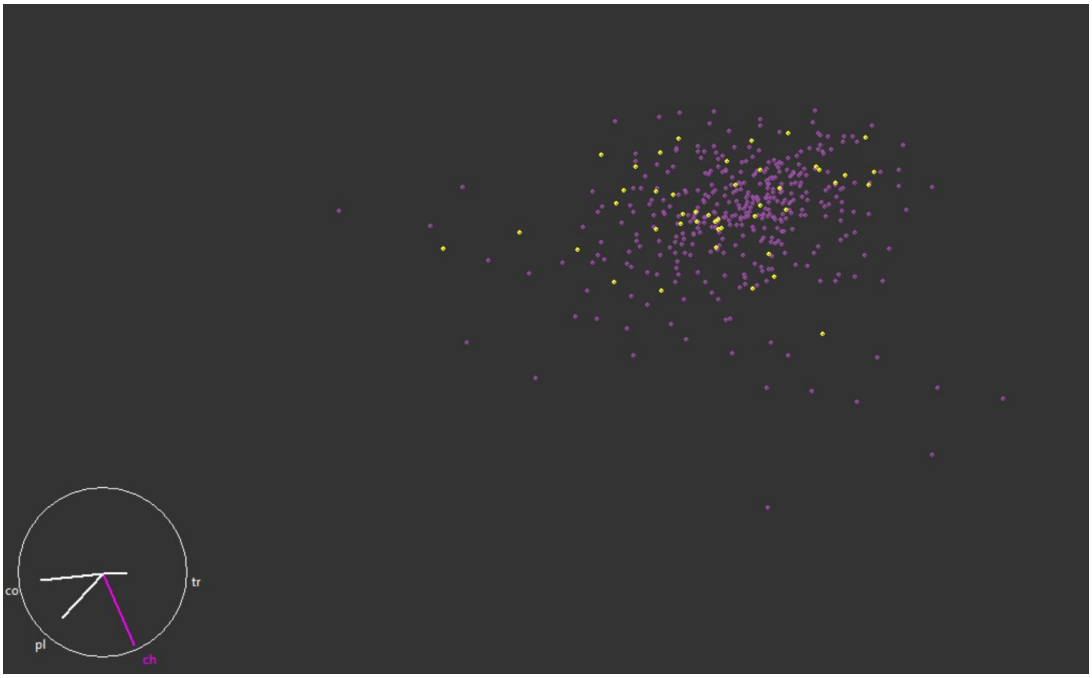
در گام بعد تور یک بعدی دو متغیر مورد نظر را باتوجه به خواسته مسئله رسم کرده و مشاهده می‌شود که توزیع توام آن‌ها تقریبا نرمال می‌باشد. (مقادیر دور افتاده هنوزهم در قسمت دُم وجود دارند)



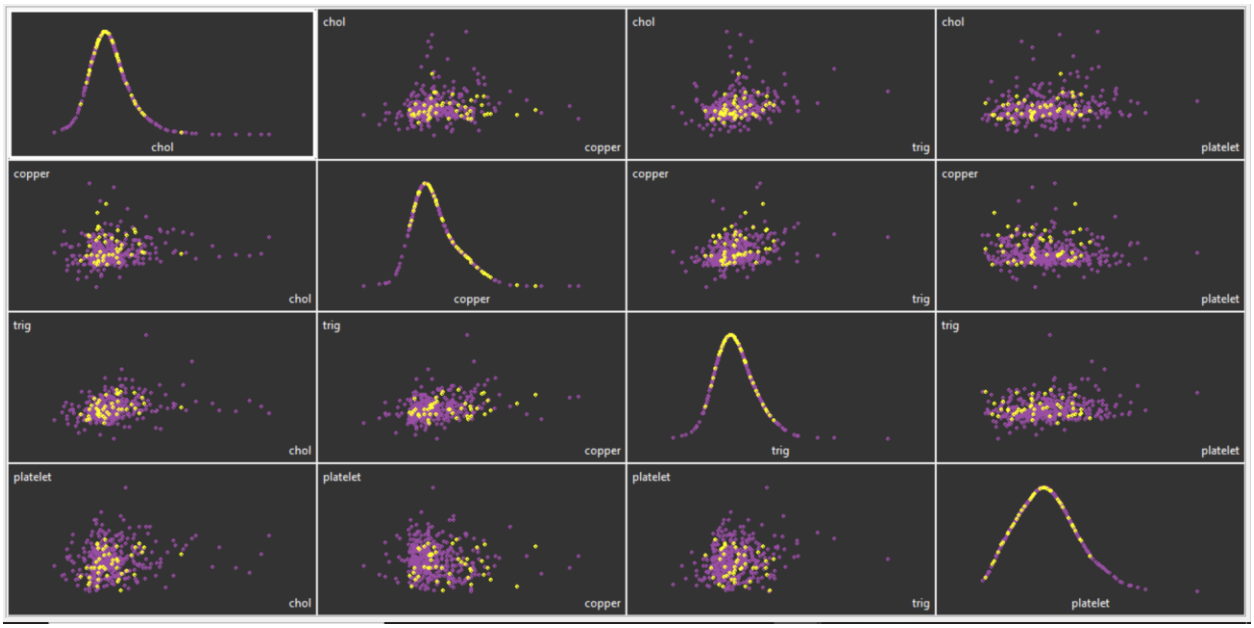
در این گام با روش multiple imputation جانهی را انجام داده و نمودار ماتریس پراکنش ۴ متغیر در نظر گرفته شده به صورت زیر می‌باشد که همانطور که مشاهده می‌شود توزیع متغیرها بیشتر از حالت اولیه، به توزیع نرمال نزدیک شده است. (همچنان روابط بین متغیرها غیر خطی است)



همچنین تور دو بعدی این متغیرها به صورت زیر است که رنگ زرد نشان دهنده گروه زن‌ها و رنگ بنفش نشان دهنده گروه مردها می‌باشد.

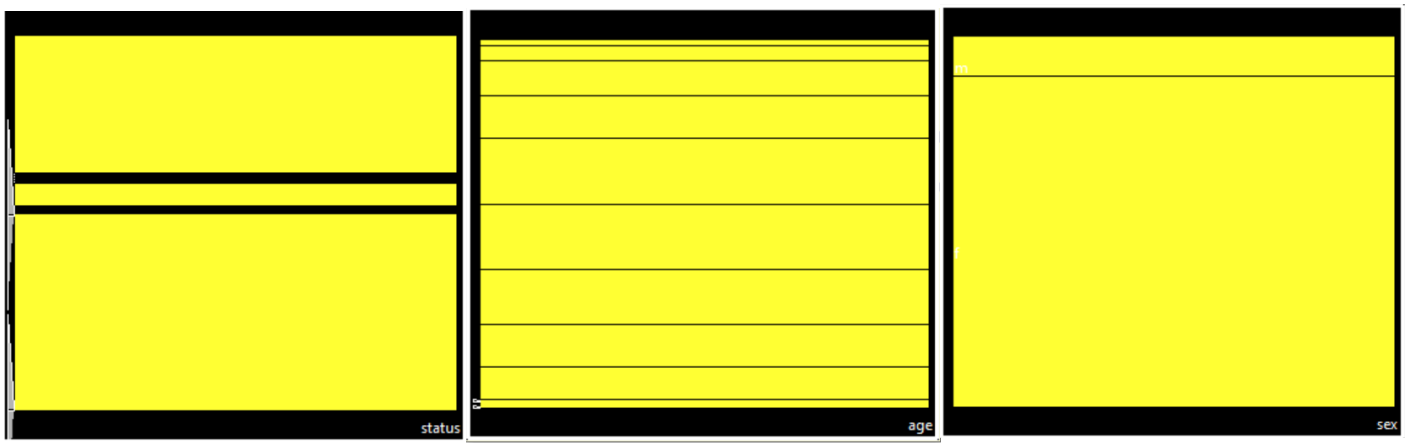


و جایگاه این گروه‌ها در نمودارهای پراکنش نیز به صورت زیر است.



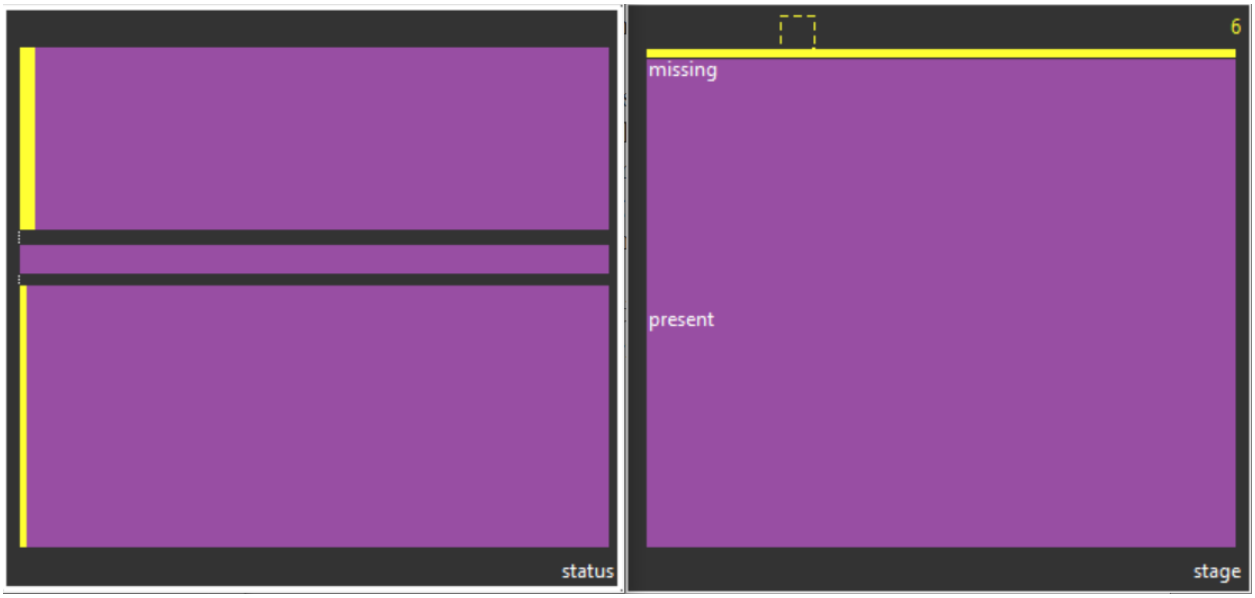
مشاهده می‌شود که بر اساس جنسیت خوشه بندی خوبی انجام نمی‌شود.

همچنین نمودارهای spine برای متغیرهای کیفی وضعیت، جنسیت و سن (در مسئله کیفی در نظر گرفته شده) به صورت زیر است.



تعداد زیادی از داده‌ها متعلق به مردان می‌باشد که اگر نمونه‌ها به صورت تصادفی انتخاب شده باشند و بیشتر روی مردان آزمایش‌ها صورت نگرفته شده باشد این یعنی، بیشتر مردها دارای این بیماری هستند. همچنین بیشترِ بیماران در سنین میان سالی می‌باشند که بیشتر در وضعیت سوم یعنی فوت شده‌ها قرار دارند و بخش کمی از آن‌ها پیوند زده شده اند.

همچنین نمودار spine plot برای مقادیر گمشده و مشاهده شده متغیر stage به صورت شکل سمت راست می‌باشد که ۶ تا از داده‌ها به رنگ زرد نمایش داده شده مربوط به داده‌های جانهی شده با استفاده از میانه است و آن مقادیر در نمودار مربوط به status در سمت چپ نیز با همان رنگ مشاهده می‌شوند.



(داده‌های جانهی شده به روش multiple imputation برای داده‌های پیوسته در فایل اکسل CleansData و داده‌های جانهی شده با استفاده از میانه برای داده‌های گسسته در فایل اکسل Cat_imputation پیوست شده است.)