

به نام خدا

تفسیر و تحلیل داده‌ها با استفاده از مصورسازی:

### تمرین (۱) رسم parallel coordinate plot برای داده‌های Flea Beetles

ابتدا داده‌ها از پکیج Flury فراخوانی و با استفاده از دستور write.table() در فایل csv ذخیره شده‌اند و بعد در ggobi بارگذاری شدند.

```
library(Flury)
```

```
write.table(Flea.Beetles, file='Flea.csv', row.names=F, sep=',')
```

همانطور که مشاهده می‌شود، داده‌ها دارای ۴ متغیر پیوسته و یک متغیر گسسته ۲ کلاسه می‌باشد.

```
> head(flea.beetles)
```

|   | Species  | TG  | Elytra | Second.Antenna | Third.Antenna |
|---|----------|-----|--------|----------------|---------------|
| 1 | oleracea | 189 | 245    | 137            | 163           |
| 2 | oleracea | 192 | 260    | 132            | 217           |
| 3 | oleracea | 217 | 276    | 141            | 192           |
| 4 | oleracea | 221 | 299    | 142            | 213           |
| 5 | oleracea | 171 | 239    | 128            | 158           |
| 6 | oleracea | 192 | 262    | 147            | 173           |

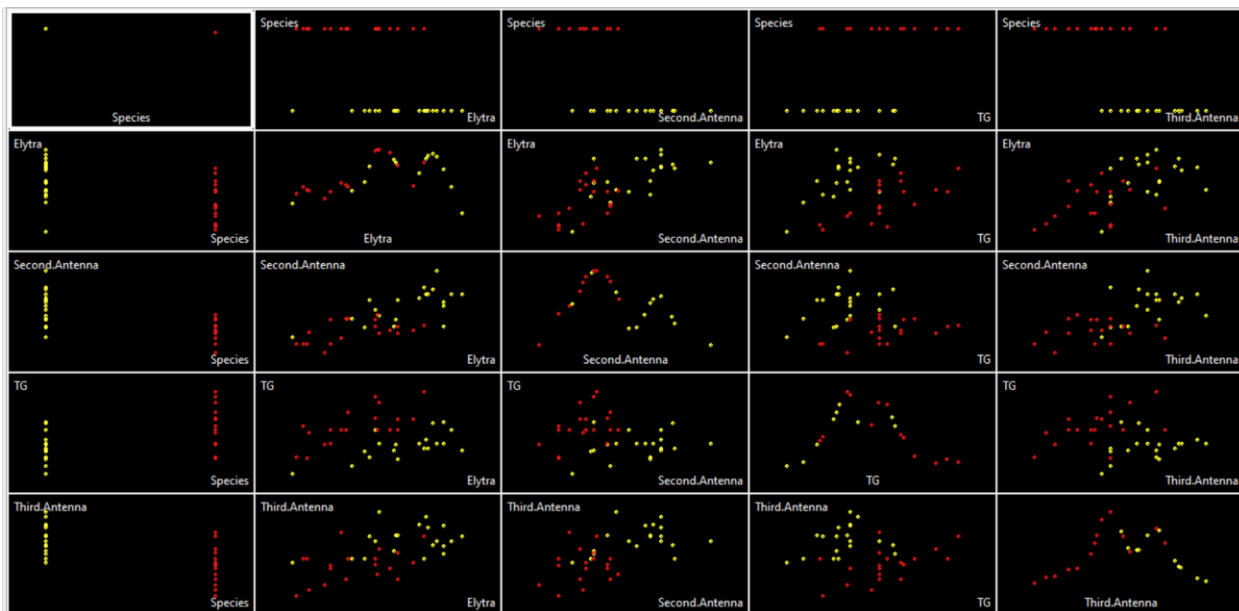
```
> summary(flea.beetles)
```

| Species      | TG            | Elytra        | Second.Antenna | Third.Antenna |
|--------------|---------------|---------------|----------------|---------------|
| oleracea :19 | Min. :158.0   | Min. :237.0   | Min. :121.0    | Min. :158.0   |
| carduorum:20 | 1st Qu.:177.0 | 1st Qu.:262.5 | 1st Qu.:137.5  | 1st Qu.:187.0 |
|              | Median :184.0 | Median :278.0 | Median :146.0  | Median :197.0 |
|              | Mean :186.8   | Mean :279.2   | Mean :147.5    | Mean :197.9   |
|              | 3rd Qu.:193.5 | 3rd Qu.:299.0 | 3rd Qu.:161.0  | 3rd Qu.:213.0 |
|              | Max. :221.0   | Max. :317.0   | Max. :184.0    | Max. :235.0   |

```
> names(flea.beetles)
```

```
[1] "Species" "TG" "Elytra" "Second.Antenna"
[5] "Third.Antenna"
```

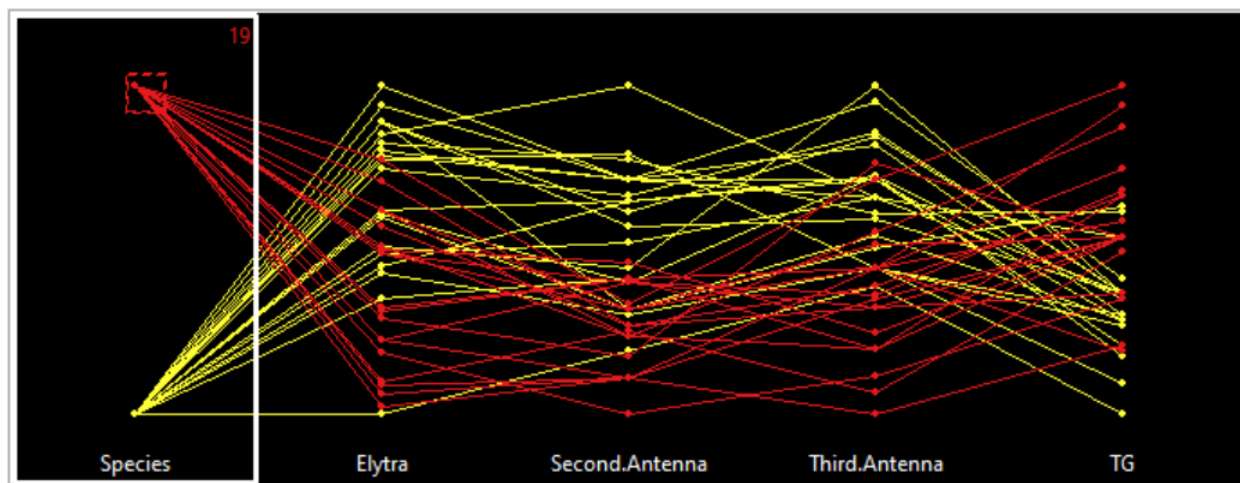
باتوجه به نمودار ماتریس پراکنش و parallel، همبستگی خطی بین متغیرها وجود ندارد.



همچنین برای متغیر Elytra و Third-Antenna مقادیر کلاس قرمز کمتر هستند اما پراکنده تر از کلاس زرد می باشند و برای متغیر Second-Antenna مقادیر قرمز همچنان کوچکتر از زرد می باشد ولی متراکمتر.

در رابطه با متغیر TG، اما، کلاس قرمز دارای مقادیر بزرگتر و پراکنده تری نسبت به کلاس زرد می باشد.

(ترتیب متغیرها جا به جا شده، و سعی شده بهترین نمودار نمایش داده شود)



سوال ۲) برای داده‌های Italian Olive Oils نمودار میله‌ای، توره‌ای ۱ بعدی، ۲ بعدی و 2x1 بعدی به شرح زیر می‌باشد.

ابتدا داده‌ها از پکیج cepp فراخوانی و با استفاده از دستور write.table() در فایل csv ذخیره شده اند و بعد در ggobi بارگذاری شدند.

```
library('cepp')
```

```
write.table(Olive, file='Olive.csv', row.names=F, sep=',')
```

داده‌های olive مربوط به روغن زیتون‌های حاوی ترکیب‌هایی از اسیدهای چرب می‌باشد که از ۳ منطقه مختلف از ایتالیا حاصل شده است.

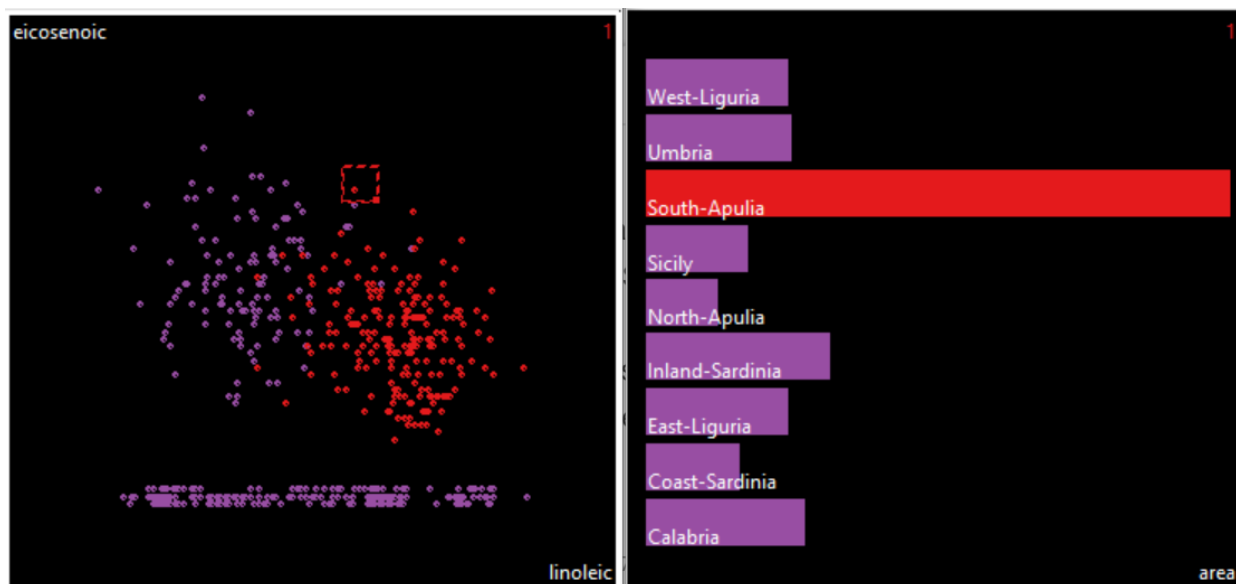
این داده‌ها شامل ۱۰ متغیر و ۵۷۲ نمونه مختلف می‌باشند (دو متغیر گسسته، یکی مربوط به ۳ ناحیه مختلف ایتالیا (شمال- جنوب- جزیره ساردینیا) و دیگری مربوط به مناطق مختلف در این ۳ ناحیه می‌باشد، همچنین ۸ متغیر پیوسته مقدار دیگر، مربوط به اسیدهای چرب مورد بحث می‌باشد) که خلاصه‌ای از داده‌ها در زیر مشاهده می‌شود.

```
> names(olive)
```

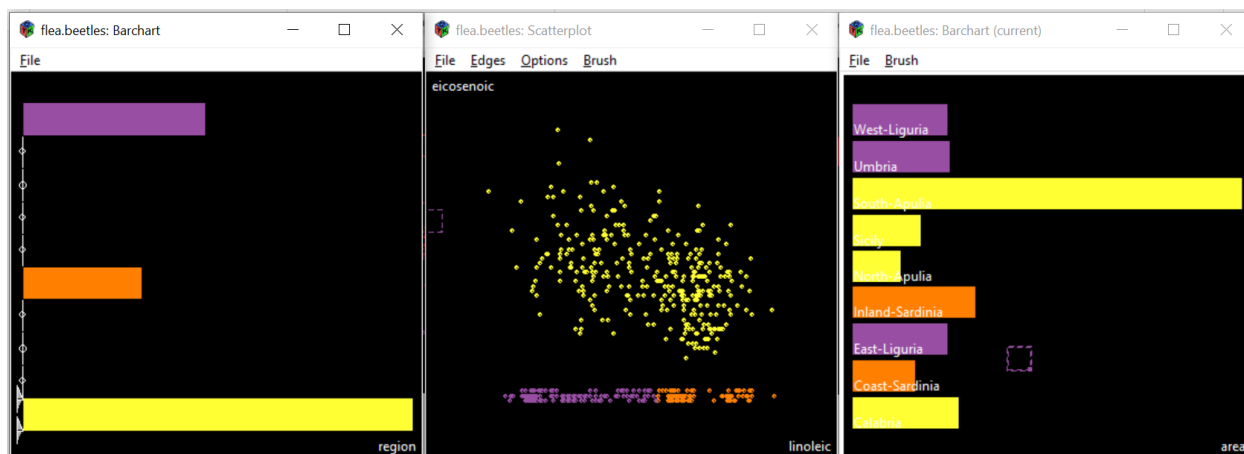
```
[1] "region"      "area"        "palmitic"    "palmitoleic" "stearic"
[6] "oleic"       "linoleic"    "linolenic"   "arachidic"   "eicosenoic"
```

| region        | area                 | palmitic       | palmitoleic    | eicosenoic    |
|---------------|----------------------|----------------|----------------|---------------|
| Min. :1.000   | South-Apulia :206    | Min. : 610     | Min. : 15.00   | Min. : 1.00   |
| 1st Qu.:1.000 | Inland-Sardinia : 65 | 1st Qu.:1095   | 1st Qu.: 87.75 | 1st Qu.: 2.00 |
| Median :1.000 | Calabria : 56        | Median :1201   | Median :110.00 | Median :17.00 |
| Mean :1.699   | Umbria : 51          | Mean :1232     | Mean :126.09   | Mean :16.28   |
| 3rd Qu.:3.000 | East-Liguria : 50    | 3rd Qu.:1360   | 3rd Qu.:169.25 | 3rd Qu.:28.00 |
| Max. :3.000   | West-Liguria : 50    | Max. :1753     | Max. :280.00   | Max. :58.00   |
|               | (Other) : 94         |                |                |               |
| stearic       | oleic                | linoleic       | linolenic      | arachidic     |
| Min. :152.0   | Min. :6300           | Min. : 448.0   | Min. : 0.00    | Min. : 0.0    |
| 1st Qu.:205.0 | 1st Qu.:7000         | 1st Qu.: 770.8 | 1st Qu.:26.00  | 1st Qu.: 50.0 |
| Median :223.0 | Median :7302         | Median :1030.0 | Median :33.00  | Median : 61.0 |
| Mean :228.9   | Mean :7312           | Mean : 980.5   | Mean :31.89    | Mean : 58.1   |
| 3rd Qu.:249.0 | 3rd Qu.:7680         | 3rd Qu.:1180.8 | 3rd Qu.:40.25  | 3rd Qu.: 70.0 |
| Max. :375.0   | Max. :8410           | Max. :1470.0   | Max. :74.00    | Max. :105.0   |

نمودار میله‌ای ناحیه‌های مختلف (area) و نمودار پراکنش متغیر eicosenoic در برابر linoleic که در آن منطقه South-Apulia با رنگ قرمز هایلایت و براش شده، به صورت زیر بدست آمده است.



در واقع نمودار سمت چپ، یک توزیع حاشیه‌ای از اسیدهای *eicosenoic* و *linoleic* را نمایش می‌دهد (نسبت به بقیه متغیرها انتگرال گرفته شده و از این قسمت حذف شده اند).



رنگ زرد مربوط به جنوب ایتالیا (ناحیه ۱)، رنگ نارنجی مربوط به جزیره ساردینیا (ناحیه ۲) و رنگ بنفش مربوط به شمال ایتالیا (ناحیه ۳) است.

باتوجه به نمودارهای بالا، مشاهده می‌شود که بیشترین تولید روغن زیتون در جنوب ایتالیا به ویژه در منطقه South-Apulia صورت گرفته است، بعد از آن بیشترین مقدار به ترتیب مربوط به ناحیه شمالی ایتالیا و جزیره ساردینیا می‌باشد.

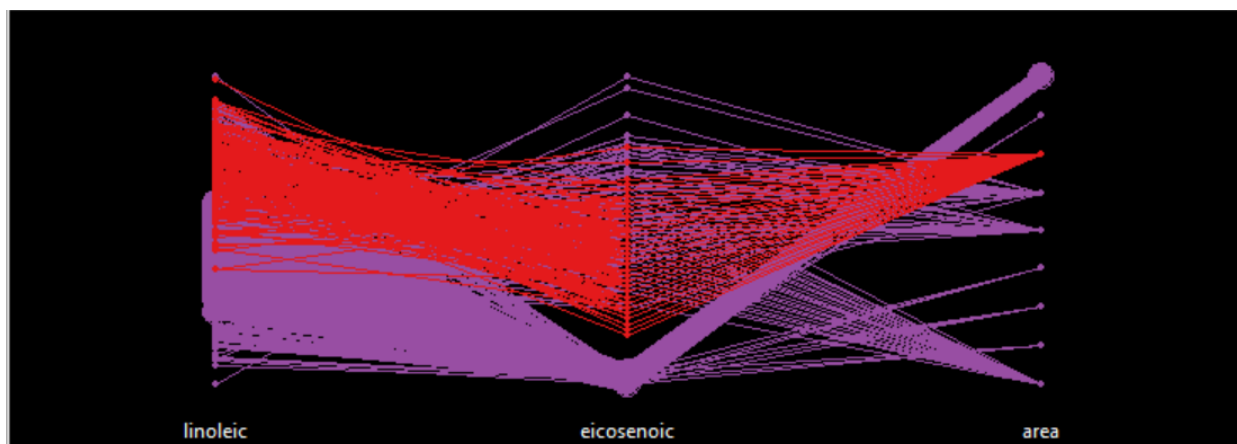
روغن‌های منطقه South-Apulia دارای اسید *linoleic* بیشتری نسبت به روغن‌های تولید شده در نواحی دیگر می‌باشند (میزان این اسید در آن‌ها برابر است با روغن‌های تولید شده در جزیره ساردینیا، پس اگر بر اساس این اسید *clustering* انجام شود این نواحی در یک *cluster* قرار می‌گیرند).

همچنین میزان اسید *eicosenoic* در روغن‌های تولید شده در ناحیه شمالی بیشتر از ۲ نواحی دیگر می‌باشد.

علاوه بر این، باتوجه به نقاط قرمز رنگ روی نمودار پراکنش به نظر می‌رسد همبستگی خطی بین این دو مقدار اسید وجود ندارد و پراکندگی شایانی بین مقادیر دیده می‌شود از این رو تفکیک کردن آن‌ها می‌تواند راحت تر باشد.

نقاط ثابتی که به ازای مقادیر مختلف اسید linoleic در پایین تصویر مشاهد می‌شوند مربوط به نواحی ۲ و ۳ می‌باشد که میزان اسیدهای linoleic و eicosenoic در آن‌ها کمتر از بقیه می‌باشد و به ازای مقادیر مختلف linoleic مقدار اسید eicosenoic ثابت باقی می‌ماند.

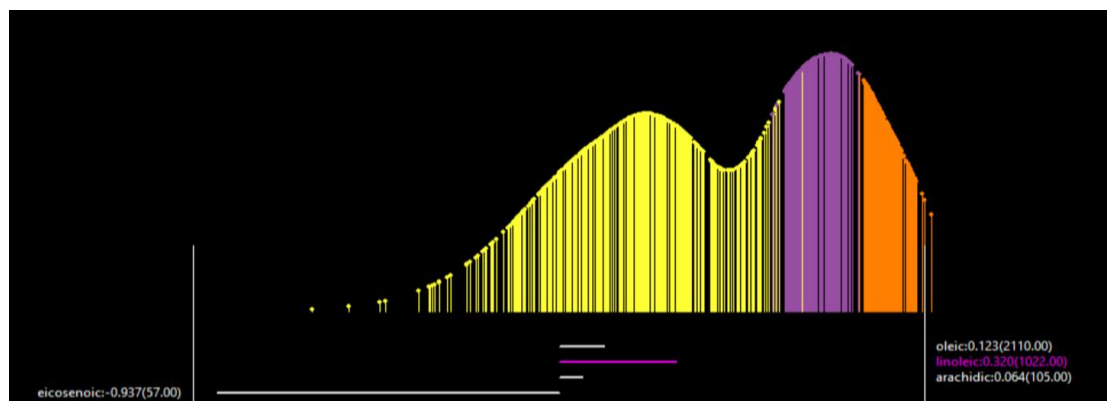
(همبستگی ضعیف بین دو متغیر بحث شده را در نمودار parallel زیر نیز می‌توان مشاهده کرد.)



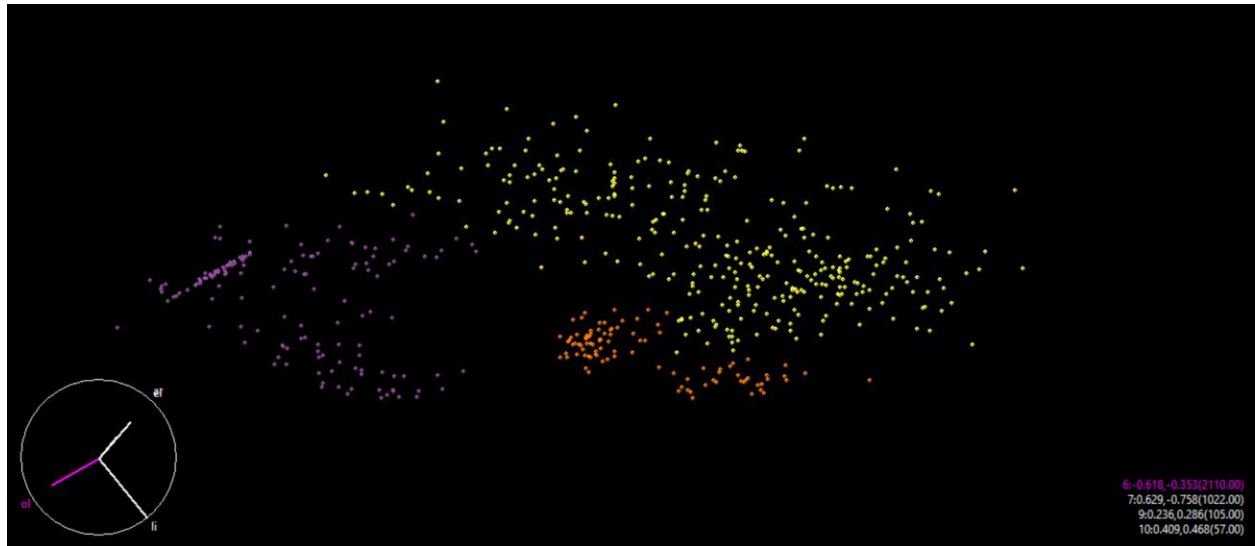
در مرحله بعد برای ۴ متغیر oleic, linoleic, arachidic و eicosenoic نمودار Grand Tour یک بعدی، دو بعدی و 2x1 بعدی رسم شده است.











با انتخاب گزینه scramble تورهای تصادفی (در واقع مولفه‌های ماتریس تصویر به صورت تصادفی انتخاب می‌شود) به وجود می‌آیند.

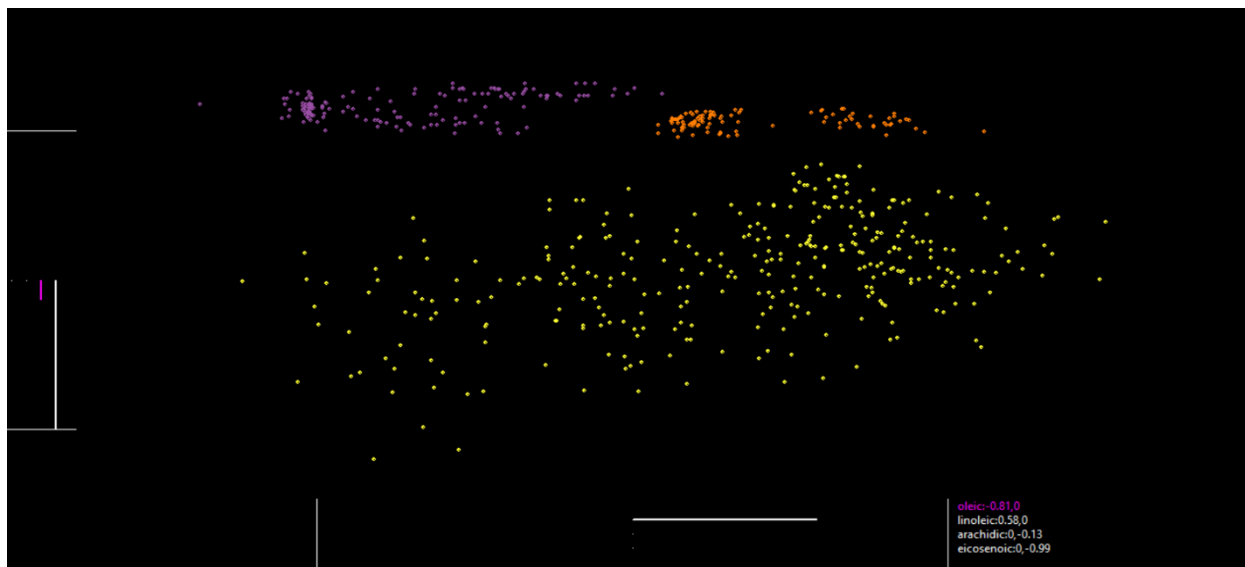
در حالت یک بعدی، با قرار دادن مقدار ASH Smoothness=0.24 و تغییر دادن ماتریس‌های تصویر به صورت تصادفی، بهترین چگالی توام بدست آمده برای این ۴ متغیر به صورت زیر می‌باشد. (بهترین در بین موارد مشاهده شده)



به صورت زیر به دست آمده است. (تفسیرها به صورت بالا می باشد و تقریبا می توان گفت تفکیک به خوبی صورت می گیرد)



|                                       |                                       |             |   |             |
|---------------------------------------|---------------------------------------|-------------|---|-------------|
| <input type="checkbox"/> X            | <input type="checkbox"/> Y            | region      |  | oleic       |
| <input type="checkbox"/> X            | <input type="checkbox"/> Y            | area        |  | linoleic    |
| <input type="checkbox"/> X            | <input type="checkbox"/> Y            | palmitic    |  | palmitic    |
| <input type="checkbox"/> X            | <input type="checkbox"/> Y            | palmitoleic |  | palmitoleic |
| <input type="checkbox"/> X            | <input type="checkbox"/> Y            | stearic     |  | stearic     |
| <input checked="" type="checkbox"/> X | <input type="checkbox"/> Y            | oleic       |  | oleic       |
| <input checked="" type="checkbox"/> X | <input type="checkbox"/> Y            | linoleic    |  | linoleic    |
| <input type="checkbox"/> X            | <input type="checkbox"/> Y            | linolenic   |  | linolenic   |
| <input type="checkbox"/> X            | <input checked="" type="checkbox"/> Y | arachidic   |  | arachidic   |
| <input type="checkbox"/> X            | <input checked="" type="checkbox"/> Y | eicosenoic  |  | eicosenoic  |



(در تمامی مراحل Brushing داده‌ها، با استفاده از گزینه Automatic brushing by variable انجام شده است.)

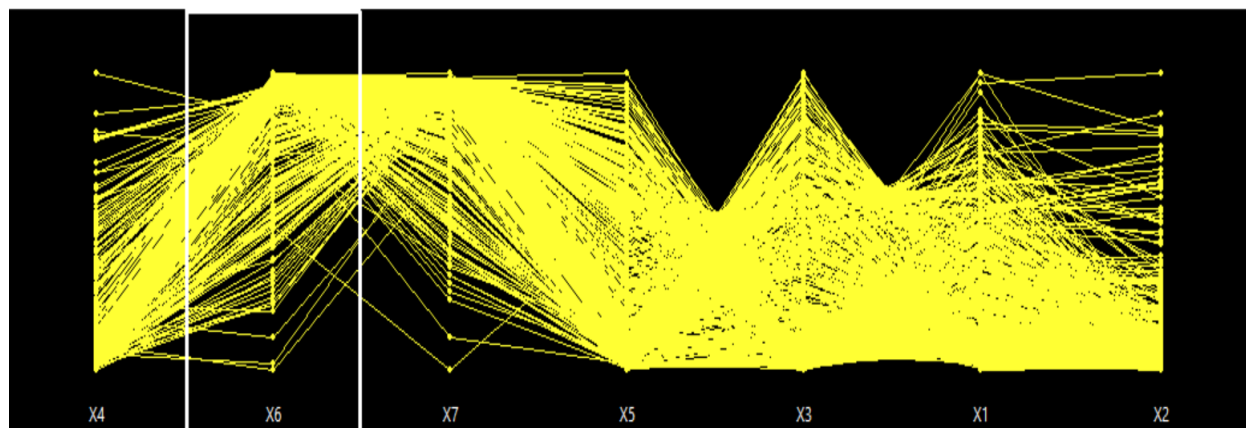
### سوال ۳) انجام sphering برای داده‌های PRIM7 و رسم Projection pursuit Tour با شاخص holes

داده‌های PRIM7 مربوط به داده‌های آزمایش فیزیک ذرات می‌باشند که دارای ۷ متغیر پیوسته و ۵۰۰ نمونه می‌باشد. این داده‌ها اولین بار در مقاله A Projection Pursuit Algorithm for Exploratory Data Analysis توسط Friedman و Tukey مورد بحث قرار گرفته اند و خلاصه‌ای از این داده‌ها به صورت زیر می‌باشد.

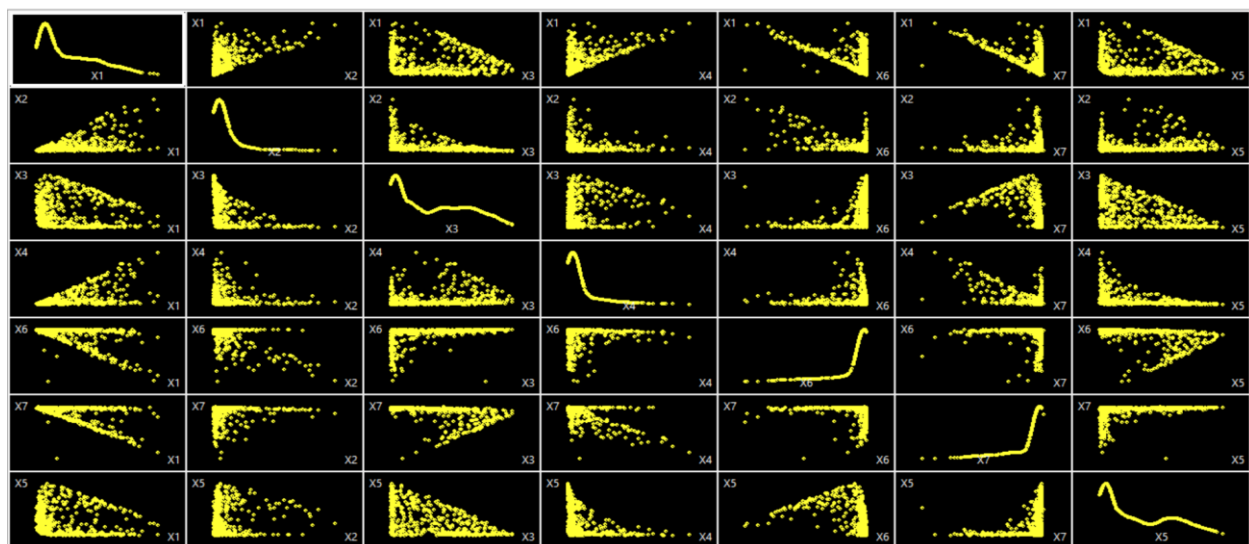
| x1               | x2              | x3               |
|------------------|-----------------|------------------|
| Min. : 0.547     | Min. : 0.0830   | Min. : 1.248     |
| 1st Qu.: 1.828   | 1st Qu.: 0.4490 | 1st Qu.: 2.317   |
| Median : 4.106   | Median : 0.7015 | Median : 8.175   |
| Mean : 5.768     | Mean : 1.5539   | Mean : 9.612     |
| 3rd Qu.: 9.110   | 3rd Qu.: 1.6772 | 3rd Qu.: 15.703  |
| Max. : 20.608    | Max. : 14.7550  | Max. : 26.124    |
| x4               | x5              | x6               |
| Min. : 0.0820    | Min. : 1.233    | Min. : -21.9000  |
| 1st Qu.: 0.4340  | 1st Qu.: 2.304  | 1st Qu.: -2.0248 |
| Median : 0.7065  | Median : 6.769  | Median : -0.6965 |
| Mean : 1.8395    | Mean : 8.860    | Mean : -2.3654   |
| 3rd Qu.: 1.7985  | 3rd Qu.: 15.168 | 3rd Qu.: -0.2120 |
| Max. : 16.7440   | Max. : 26.417   | Max. : -0.0140   |
| x7               |                 |                  |
| Min. : -24.2450  |                 |                  |
| 1st Qu.: -2.3708 |                 |                  |
| Median : -0.6905 |                 |                  |
| Mean : -2.5324   |                 |                  |
| 3rd Qu.: -0.2037 |                 |                  |
| Max. : 0.3880    |                 |                  |

داده‌های PRIM7 در کتابخانه groc قابل دسترس می‌باشند.

باتوجه به نمودار parallel coordinate زیر به نظر می‌رسد بین بعضی متغیرها مثل  $X_1$  و  $X_2$  همبستگی وجود دارد پس نیازی نیست از همه آن‌ها در تحلیل و مدل سازی استفاده کرد.



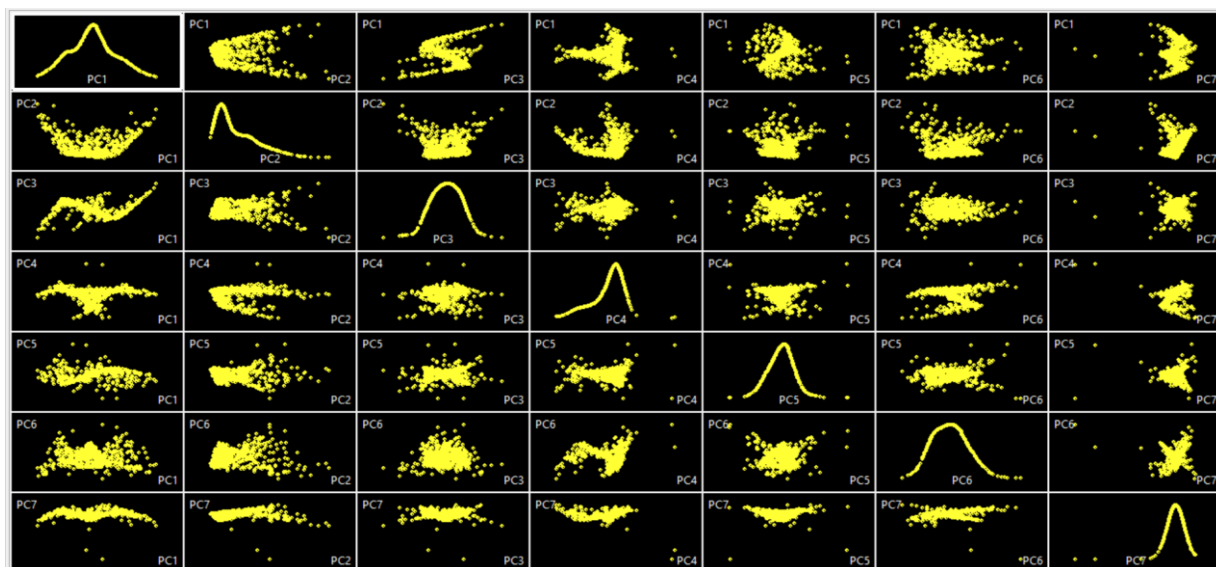
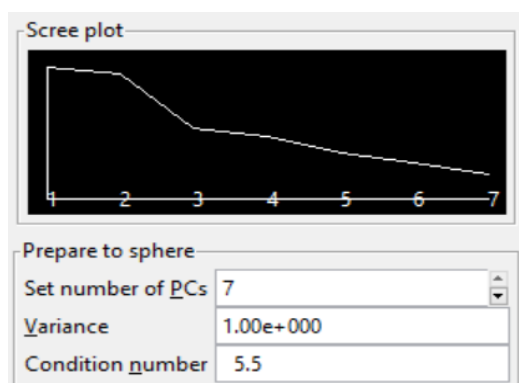
همچنین نمودار ماتریس پراکنش در یافتن همبستگی بین متغیرها به ما کمک می‌کند.



همانطور که مشاهده می‌شود بین متغیرها همبستگی خطی قوی وجود ندارد.

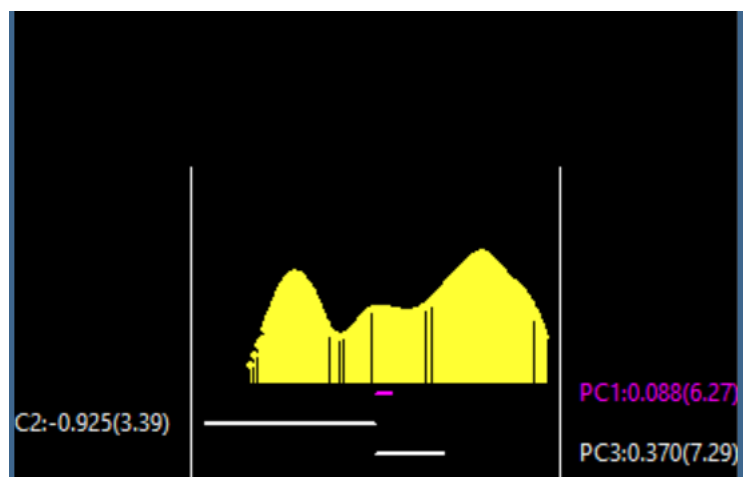


اما با انجام sphering روی داده‌ها، با توجه به نمودار Scree plot اولین شکستگی و زانو در نقطه ۳ مشاهده می‌شود که به نظر می‌رسد با استفاده از سه متغیر به میزان مورد نظر برای واریانس دست یافته ایم.

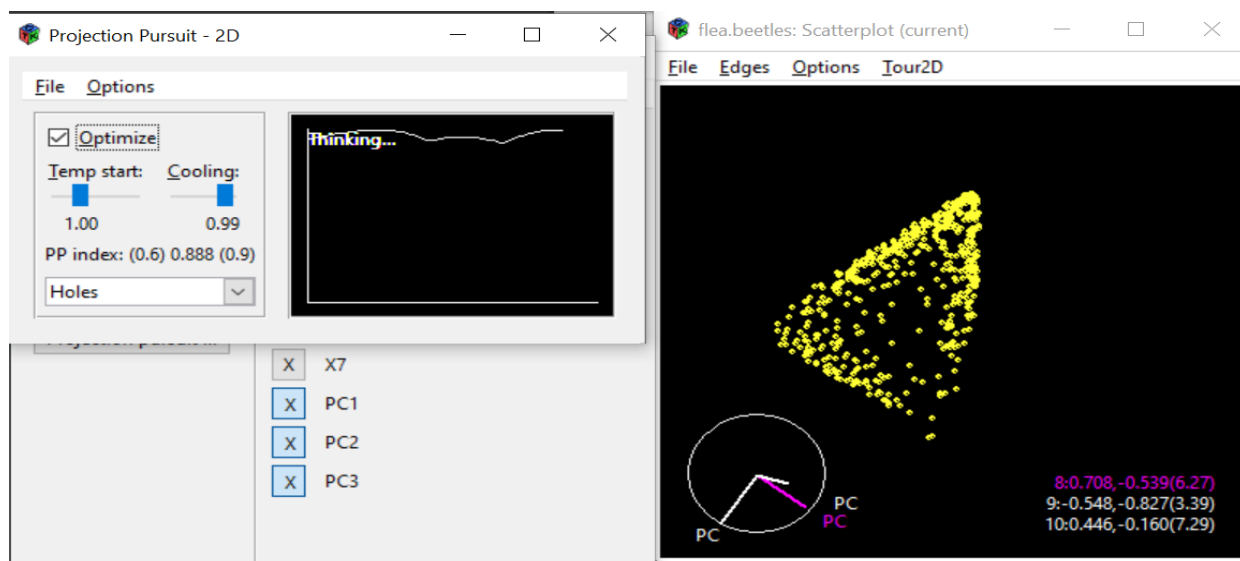


با توجه به نمودار بالا، همبستگی بین PC1 و PC5 و PC1 و PC6 و PC4 و PC3 وجود ندارد.

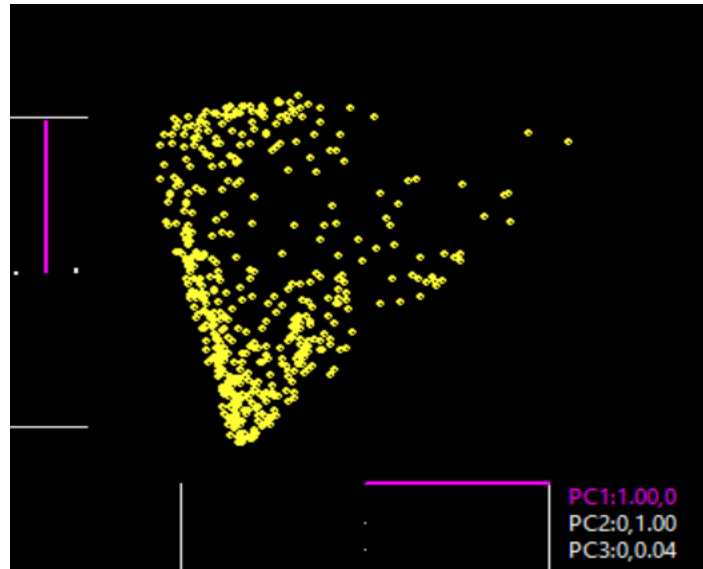
و در ادامه با استفاده از مولفه‌های اصلی یافت شده و به ترتیب تورهای ۱ بعدی، دو بعدی و 2x1 بعدی که با استفاده از شاخص holes پدید آمده است.



تور یک بعدی بالا نمایی از تابع چگالی توام سه مولفه اصلی بدست آمده، می‌باشد. همانطور که مشاهده می‌شود، نمودار چند مَدی می‌باشد و واریانس بیشتر (از حالت یک متغیره) شده که این نشان دهنده آن است که این داده‌ها تفکیک پذیرترند و همبستگی آن‌ها کم شده است. همچنین تور دو بعدی و 2x1 بعدی حاصل از این داده‌ها به ترتیب به صورت زیر می‌باشد که نمایی از این سه متغیر را در فضای دو بعدی نمایش می‌دهند.



2D Tour



2x1 D Tour

(در اینجا برای رسم 2x1 D Tour متغیر PC1 به عنوان x و متغیرهای PC2 و PC3 به عنوان y در نظر گرفته شده اند.)

برای نمایش حالت interactive نمودارها، انیمشن‌هایی از Tourها در فایل‌های gift قرار داده شده است.

#### سوال ۴) استفاده از identify برای داده‌های wages

(به نظر می‌رسد منظور مسئله، داده‌های Wage باشد به خاطر اینکه اطلاعات راجع به زمان در اینجا قرار دارد.)

این داده‌ها مربوط به دستمزد گروهی از کارگران مرد منطقه میانی اقیانوس اطلس می‌باشند و شامل ۱۱ متغیر و ۳۰۰۰ نمونه است که از کتابخانه ISLR قابل دسترس هستند.

بخشی از این داده‌ها و خلاصه‌ای از آن در زیر، نمایش داده شده است.

```
> head(wage)
  year age   maritl   race   education   region   jobclass
231655 2006  18 1. Never Married 1. white  1. < HS Grad 2. Middle Atlantic 1. Industrial
86582  2004  24 1. Never Married 1. white  4. College Grad 2. Middle Atlantic 2. Information
161300 2003  45   2. Married 1. white  3. Some College 2. Middle Atlantic 1. Industrial
155159 2003  43   2. Married 3. Asian  4. College Grad 2. Middle Atlantic 2. Information
11443  2005  50   4. Divorced 1. white  2. HS Grad 2. Middle Atlantic 2. Information
376662 2008  54   2. Married 1. white  4. College Grad 2. Middle Atlantic 2. Information

  health health_ins logwage   wage
231655  1. <=Good    2. No 4.318063 75.04315
86582  2. >=Very Good 2. No 4.255273 70.47602
161300  1. <=Good    1. Yes 4.875061 130.98218
155159 2. >=Very Good 1. Yes 5.041393 154.68529
11443  1. <=Good    1. Yes 4.318063 75.04315
376662 2. >=Very Good 1. Yes 4.845098 127.11574
```

```
> summary(wage)
```

| year    |       | age     |        | maritl            |       | race      |      | education           |      |
|---------|-------|---------|--------|-------------------|-------|-----------|------|---------------------|------|
| Min.    | :2003 | Min.    | :18.00 | 1. Never Married: | 648   | 1. White: | 2480 | 1. < HS Grad        | :268 |
| 1st Qu. | :2004 | 1st Qu. | :33.75 | 2. Married        | :2074 | 2. Black: | 293  | 2. HS Grad          | :971 |
| Median  | :2006 | Median  | :42.00 | 3. Widowed        | : 19  | 3. Asian: | 190  | 3. Some College     | :650 |
| Mean    | :2006 | Mean    | :42.41 | 4. Divorced       | : 204 | 4. Other: | 37   | 4. College Grad     | :685 |
| 3rd Qu. | :2008 | 3rd Qu. | :51.00 | 5. Separated      | : 55  |           |      | 5. Advanced Degree: | 426  |
| Max.    | :2009 | Max.    | :80.00 |                   |       |           |      |                     |      |

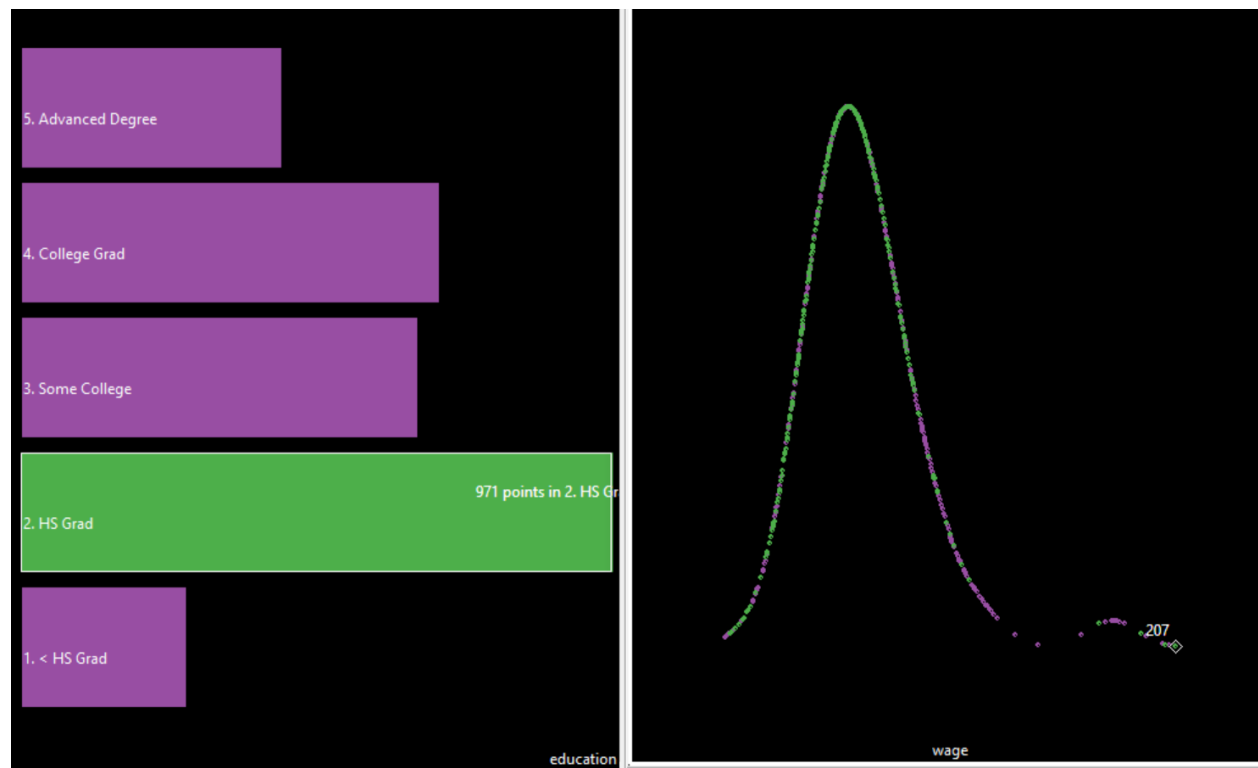
  

| region                 |       | jobclass       |       | health         |       | health_ins |      |
|------------------------|-------|----------------|-------|----------------|-------|------------|------|
| 2. Middle Atlantic     | :3000 | 1. Industrial  | :1544 | 1. <=Good      | : 858 | 1. Yes:    | 2083 |
| 1. New England         | : 0   | 2. Information | :1456 | 2. >=Very Good | :2142 | 2. No :    | 917  |
| 3. East North Central: | 0     |                |       |                |       |            |      |
| 4. West North Central: | 0     |                |       |                |       |            |      |
| 5. South Atlantic      | : 0   |                |       |                |       |            |      |
| 6. East South Central: | 0     |                |       |                |       |            |      |
| (Other)                | : 0   |                |       |                |       |            |      |

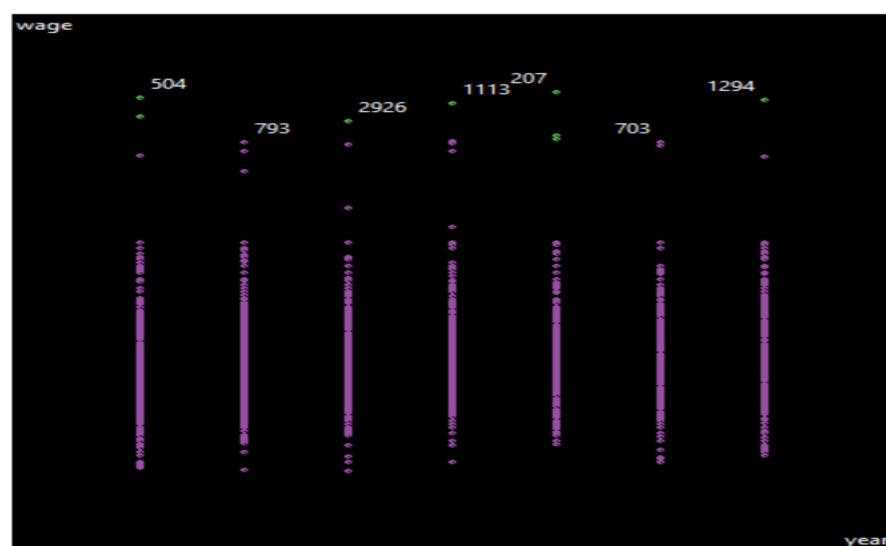
| logwage |        | wage    |         |
|---------|--------|---------|---------|
| Min.    | :3.000 | Min.    | : 20.09 |
| 1st Qu. | :4.447 | 1st Qu. | : 85.38 |
| Median  | :4.653 | Median  | :104.92 |
| Mean    | :4.654 | Mean    | :111.70 |
| 3rd Qu. | :4.857 | 3rd Qu. | :128.68 |
| Max.    | :5.763 | Max.    | :318.34 |

باتوجه به شکل زیر، در می‌یابیم که بیشتر کارگرا دارای مدرک دیپلم می‌باشند و بیشترین دستمزد برای آن‌ها حدودا بین ۸۵ تا ۱۰۵ دلار می‌باشد.

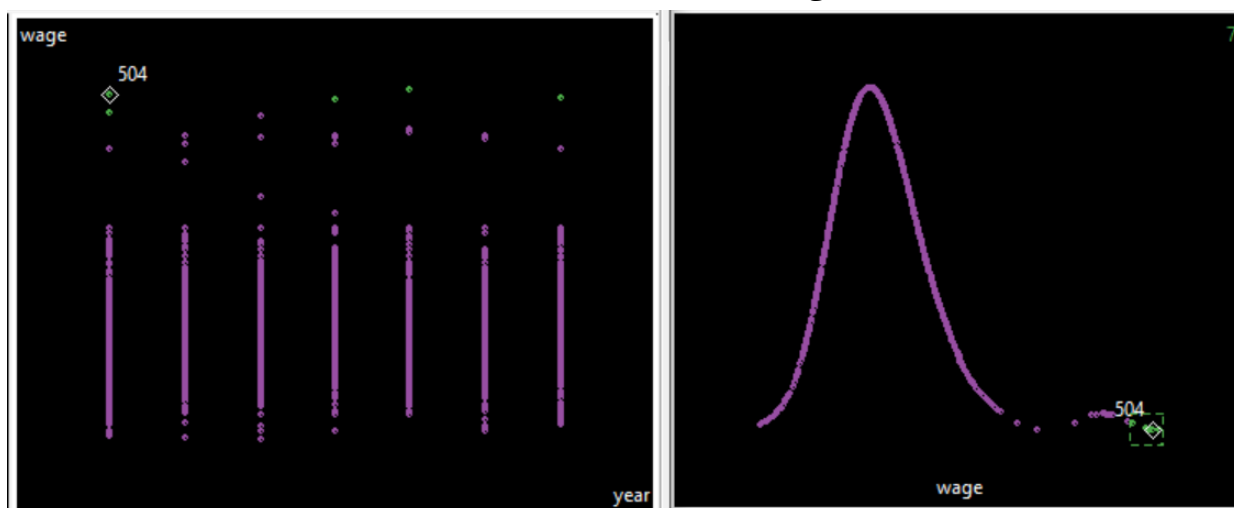


همچنین فرد 207ام با مدرک دیپلم دارای بیشترین دستمزد می‌باشد.

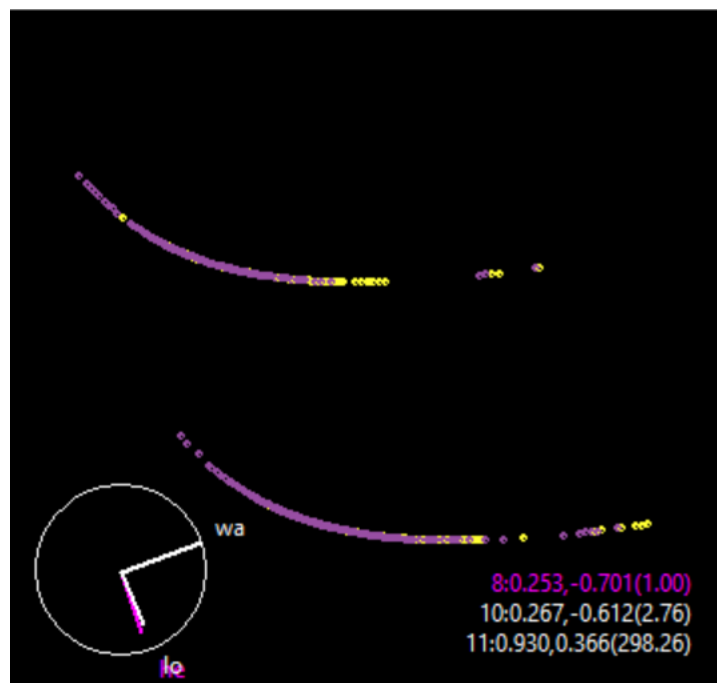
نمونه‌های سبز رنگ در شکل زیر، مربوط به افرادی است که در هر سال دارای بیشترین دستمزد می‌باشند یعنی افراد ۱۲۹۴م، ۷۰۳م، ۲۰۷م، ۱۱۱۳م، ۲۹۲۶م، ۷۹۳م و ۵۰۴م.



که در بین افرادی که در سال ۲۰۰۹ دارای درآمد بالایی هستند، فرد ۵۰۴م از ۶ سال قبل هم (سال ۲۰۰۳-اولین سال بیان شده در مجموعه داده‌ها) دارای بالاترین دستمزد می‌باشد.



همچنین یک تور دو بعدی متشکل از متغیرهای دستمزد، وضعیت سلامت و  $\log(wage)$  با brushing متغیر وضعیت بیمه به صورت زیر می‌باشد.



همانطور که مشاهده می‌شود تفکیک این متغیرها به وسیله متغیر وضعیت بیمه، به خوبی انجام نشده اما به وسیله متغیر وضعیت سلامت که در شکل زیر نمایش داده شده تفکیک دو گروه (وضعیت سلامت تقریباً کم - وضعیت سلامت تقریباً کامل و خوب) به خوبی صورت گرفته است.

