به نام خدا


گزارش فاز اول پروژه مقدمه‌ای بر بیوانفورماتیک

استاد

**دکتر شریفی و کوهی**

نویسندگان

**امیرمحمد ایمانی، سیاوش رحیمی شاطرانلو، امیرحسین عابدی**

**( به ترتیب الفبا )**

## محتواها


۱ Micro-Array ..... ۳

۲ بررسی کیفیت داده‌ها ..... ۴

۳ کاهش ابعاد ..... ۶

۱.۳ UMAP ..... ۶

۲.۳ PCA ..... ۹

3.3 TSNE ..... ۱۱

۴ تحلیل نمونه‌ها ..... ۱۳

## ۱ Micro-Array

در ابتدا لازم است درمورد داده‌های بدست آمده از micro-array توضیح دهیم.

Micro-array یک لوازم آزمایشگاهی است که با استفاده از آن میتوانیم میزان بیان تعداد زیادی از ژن‌ها را برای یک DNA خاص مشخص کنیم. این لوازم صفحاتی دارند که با قرار گرفتن cDNA بر روی این صفحات مقدار بیان یک ژن در آنها مشخص میشود.

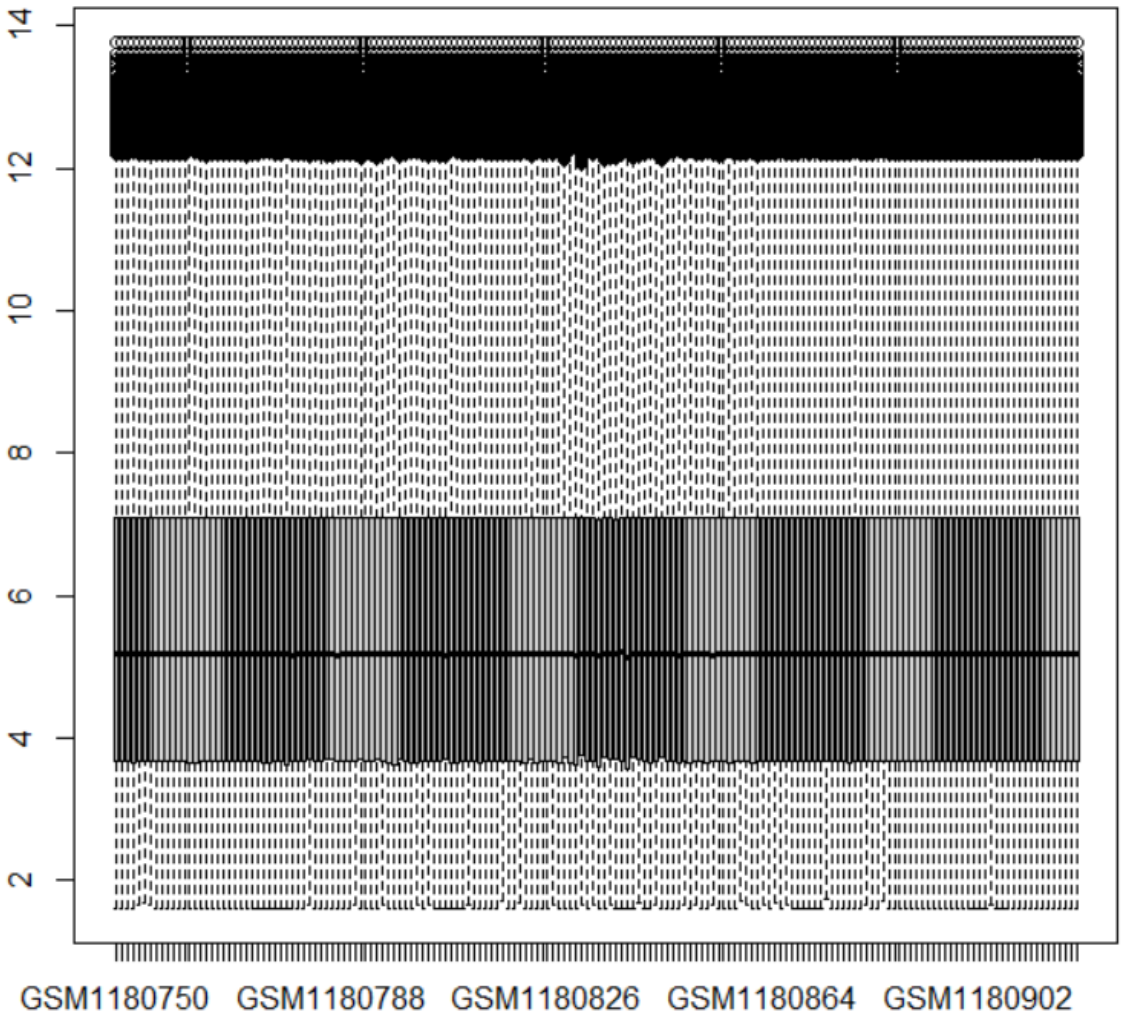
اینکار را برای چند سمپل انجام میدهیم. به طور مثال انسان‌هایی که سالم هستند و ( در این مسئله ) نمونه‌هایی که سرطان AML دارند.

داده خروجی آن که قرار است در ادامه بر روی آن کار کنیم به صورت یک ماتریس است که آنرا در شکل زیر میبینیم :

| | GSM1180750 | GSM1180751 | GSM1180752 | GSM1180753 | GSM1180754 | GSM1180755 | GSM1180756 | GSM1180757 | GSM1180758 |
|---|---|---|---|---|---|---|---|---|---|
| 7892501 | 5.635547 | 4.916813 | 5.478152 | 5.596580 | 5.768478 | 6.847387 | 3.805093 | 6.656674 | 5.699819 |
| 7892502 | 6.640414 | 5.838517 | 7.101921 | 5.190309 | 5.926461 | 7.883791 | 5.893643 | 6.656937 | 5.397055 |
| 7892503 | 5.108161 | 5.953453 | 6.383739 | 3.696127 | 5.701286 | 5.718447 | 4.426680 | 4.329873 | 3.767084 |
| 7892504 | 8.414047 | 9.026401 | 9.456269 | 8.746534 | 7.717569 | 7.526121 | 8.438070 | 7.896443 | 8.290602 |
| 7892505 | 2.280691 | 2.423883 | 3.141614 | 2.105208 | 3.035931 | 3.692030 | 3.174492 | 2.914909 | 2.977506 |
| 7892506 | 4.123770 | 5.783080 | 4.898264 | 5.199521 | 6.339047 | 6.029651 | 5.778521 | 5.361175 | 6.290925 |
| 7892507 | 3.894260 | 3.538837 | 3.589695 | 3.438865 | 2.932658 | 4.380369 | 4.998056 | 4.517858 | 4.249883 |
| 7892508 | 6.138993 | 6.827082 | 6.837599 | 5.653893 | 7.524076 | 7.649780 | 4.696382 | 6.850862 | 5.729190 |
| 7892509 | 10.743463 | 10.374949 | 11.340184 | 10.623438 | 11.094760 | 11.094760 | 11.366706 | 11.190767 | 11.280896 |
| 7892510 | 3.692390 | 4.667512 | 6.292694 | 3.480283 | 4.768568 | 7.097245 | 5.128661 | 5.840359 | 3.745148 |
| 7892511 | 2.588934 | 4.226223 | 3.257911 | 2.603042 | 2.919893 | 4.275954 | 3.853780 | 3.175632 | 2.357797 |
| 7892512 | 6.841266 | 6.023310 | 6.205062 | 6.311448 | 6.828303 | 5.923473 | 6.681854 | 5.957436 | 6.209187 |
| 7892513 | 2.639889 | 3.538837 | 3.891498 | 3.211589 | 2.811503 | 5.207296 | 4.404617 | 4.685398 | 2.393750 |
| 7892514 | 10.343113 | 9.715963 | 10.053727 | 10.212248 | 10.204758 | 10.383443 | 8.930210 | 9.868466 | 9.428465 |
| 7892515 | 8.018960 | 7.848879 | 8.586365 | 7.819625 | 7.562850 | 7.576732 | 7.815356 | 7.102698 | 7.682163 |
| 7892516 | 4.796953 | 4.102431 | 5.697539 | 3.290746 | 5.853061 | 3.565581 | 5.505044 | 5.386854 | 5.168036 |
| 7892517 | 7.087301 | 7.289988 | 6.477558 | 5.326537 | 7.243589 | 7.931539 | 6.308732 | 6.924349 | 7.640763 |
| 7892518 | 2.336036 | 2.769308 | 2.916904 | 2.521251 | 2.542003 | 2.868407 | 2.336996 | 4.054067 | 2.741425 |
| 7892519 | 5.719596 | 5.803996 | 6.341742 | 4.588688 | 5.498574 | 5.775285 | 5.331192 | 6.912587 | 5.165096 |

## ۲ بررسی کیفیت داده‌ها

در ابتدا چندمعیار برای درست بودن داده‌هایی که در دست داریم چک میکنیم. در ابتدا میتوانیم از خروجی boxplot برای این داده‌ها استفاده کنیم.




همانطور که مشاهده میشود خط مشکی وسط برای تمامی نمونه‌ها یکی میباشد. یعنی میانه میزان بیان ژن‌ها برای تمامی آنها به طور تقریبی برابر میشود. این یکی از معیارهایی است که

به ما نشان میدهد که داده‌هایی که در دست داریم داده‌های خوبی هستند و نیاز به Augment کردن آنها نیست. در صورتی که این اتفاق نمی‌افتد میتوانستیم نتیجه بگیریم که داده‌هایی که بدست آوردیم میتواند اشتباه باشد و DNA هایی که داریم مربوط به نمونه‌های ما نیستند.

یکی دیگر از روش‌هایی که میتوانیم از صحت داده‌ها ( و تست ) مطمئن شویم انجام آزمون‌های فرض بر روی داده‌ها است. توجه کنید که از بین ژن هایی که میزان بیان آنهارا داریم حداقل میزان بیان یکی از آنها باید بر روی نمونه‌های مختلف اثر گذاشته باشد. اگر اینطور نباشد داده‌ها درست نیستند و جایی در روند استخراج داده‌ها اشتباهی رخ داده است.

برای استناد به تاثیرگذاری میزان بیان ژن‌ها از آزمون فرض استفاده میکنیم و از روی مقدار p.val ها میتوانیم به درستی فرض خود پی ببریم. فرض کنید در این قسمت داشته باشیم $\alpha = 0.01$.

با استفاده از روش Limma میتوانیم به adj.p.val ها دسترسی داشته باشیم. قسمتی از این جدول به شرح زیر است.

| logFC | AveExpr | t | P.Value | adj.P.Val | B |
|---|---|---|---|---|---|
| 4.2412917 | 3.953450 | 16.445385 | 4.953819e-37 | 1.601124e-32 | 72.22863 |
| 3.3168649 | 4.807230 | 12.067697 | 1.215355e-24 | 1.964074e-20 | 44.93101 |
| 4.3928712 | 5.393336 | 11.640885 | 1.987080e-23 | 2.140814e-19 | 42.24251 |
| 3.7546691 | 3.922798 | 11.555688 | 3.467086e-23 | 2.801492e-19 | 41.70671 |
| 3.9847926 | 4.088602 | 11.354090 | 1.291853e-22 | 8.350797e-19 | 40.44034 |
| 2.1725465 | 4.377762 | 11.067976 | 8.312530e-22 | 4.477821e-18 | 38.64738 |
| 3.2456570 | 3.575768 | 10.952165 | 1.762573e-21 | 8.138304e-18 | 37.92334 |
| 3.2710142 | 3.307306 | 10.919714 | 2.175280e-21 | 8.788404e-18 | 37.72066 |
| 3.2881319 | 5.087973 | 10.361231 | 7.985842e-20 | 2.867893e-16 | 34.24835 |
| 5.3359659 | 5.491700 | 10.224293 | 1.920852e-19 | 5.728571e-16 | 33.40228 |
| 3.8078252 | 3.951406 | 10.221969 | 1.949639e-19 | 5.728571e-16 | 33.38794 |

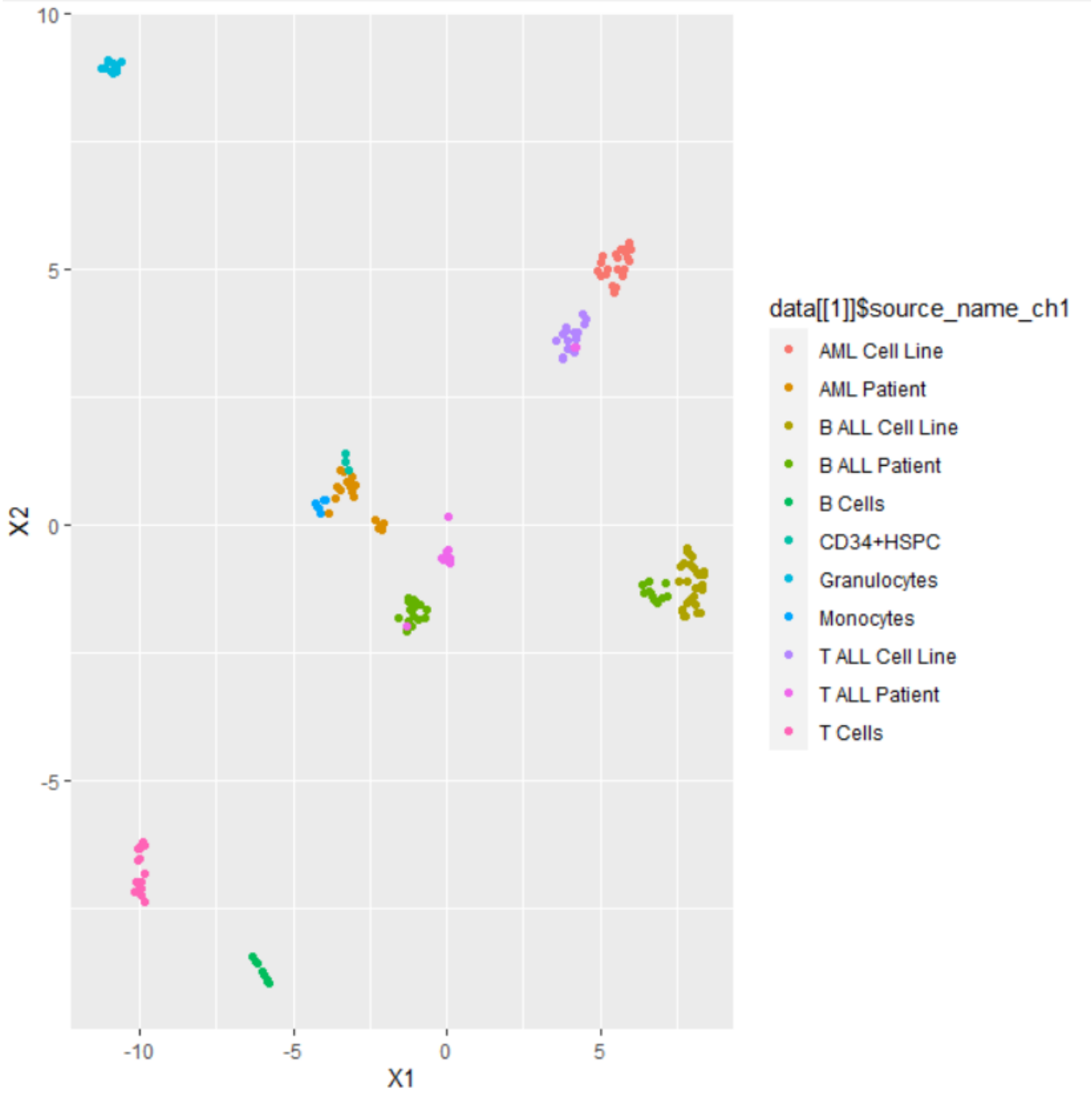
بنابراین به نظر میرسد که قسمتی از ژن ها وجود دارند که میزان بیان آنها تاثیری معنادار ایجاد کرده است. و داده‌ها کیفیت موردنظر را دارند.

## ۳ کاهش ابعاد

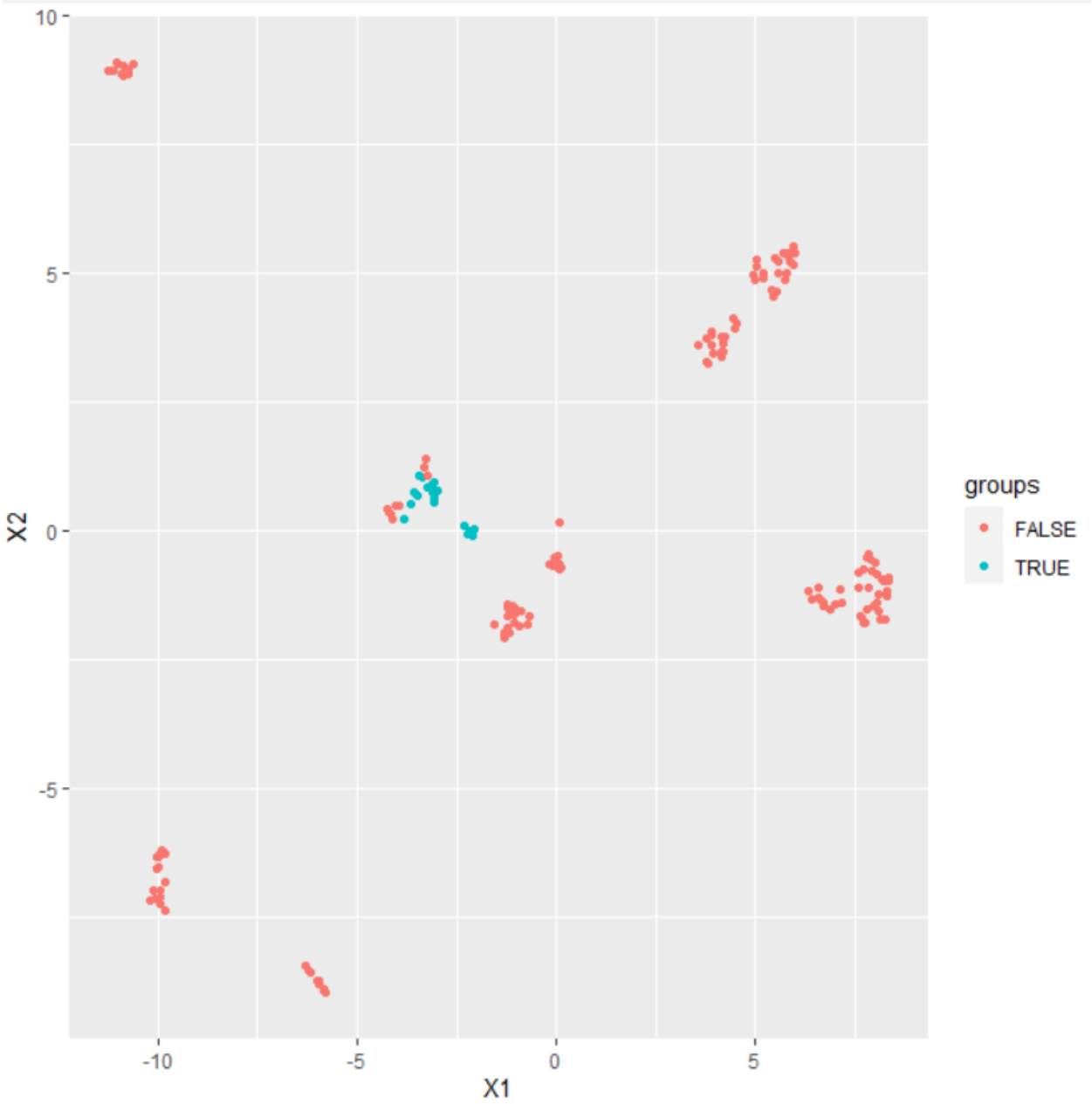
یکی از روش‌های دیگری که میتوانیم از صحت تست خود اطمینان حاصل کنیم این است که داده‌های خود را به بعد ۲ ببریم و پراکندگی آنهارا چک کنیم. تستی که درست انجام شده باشد باید به درستی نمونه‌های مختلف را از هم جدا کند. در اینجا از روش های UMAP-PCA-TSNE استفاده میکنیم.

### 1.3 UMAP

در ابتدا از روش UMAP استفاده میکنیم. این الگوریتم داده‌ها را به صورت زیر بر روی 2 بعد نمایش میدهد :




که در این مورد رنگ‌های مختلف مربوط به source name های مختلف هستند. اگر بخواهیم هرکدام از داده‌ها را با رنگ سالم و ناسالم نشان دهیم به نمودار زیر میرسیم.

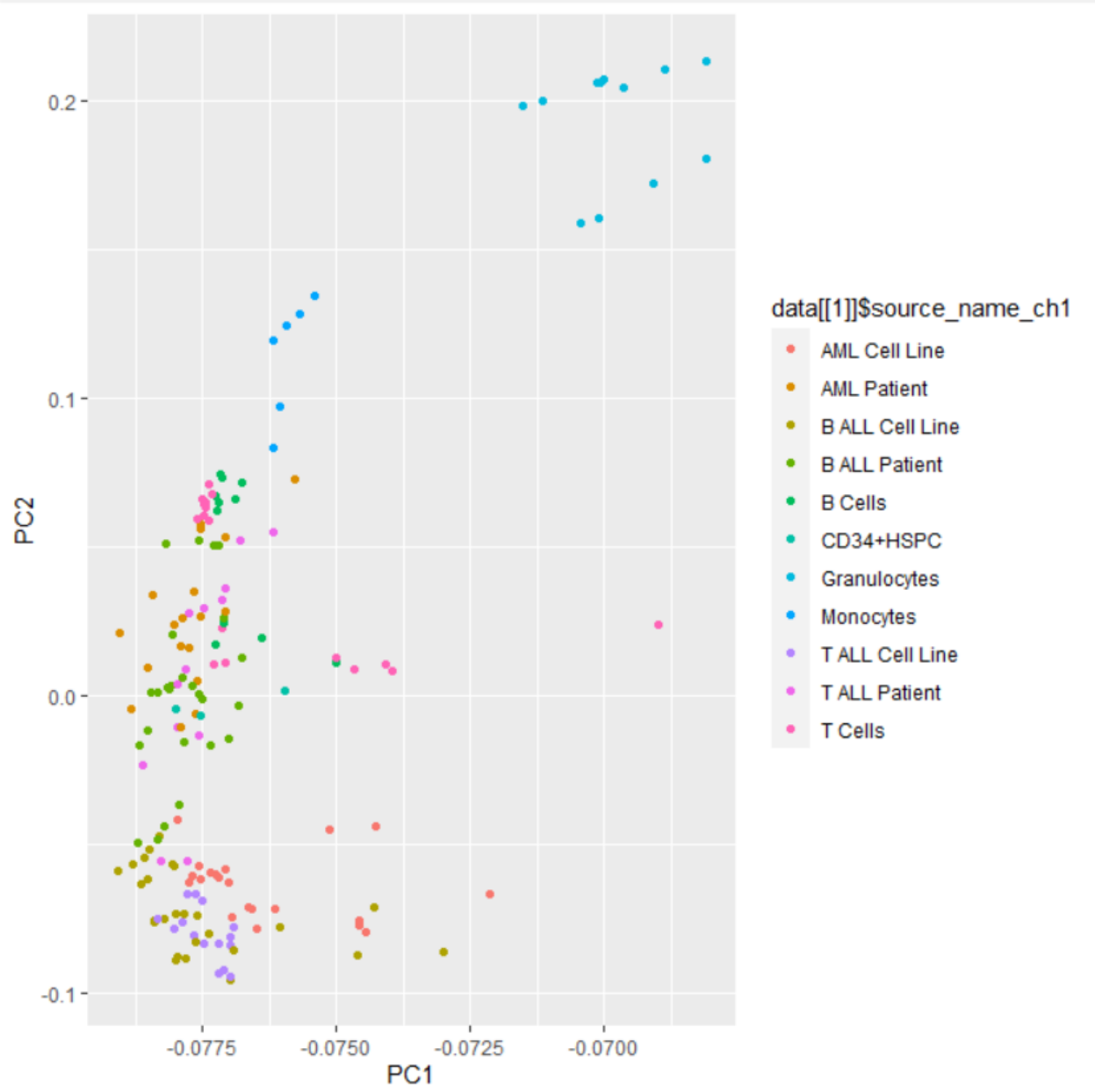



داده‌های آبی مربوط به بیماران است و بنظر میرسد که داده‌ها کیفیت خوبی دارند. ( دقت کنید که در این نمودار گروهی از افراد سالم وجود دارند که بسیار به افراد ناسالم نزدیک هستند. )

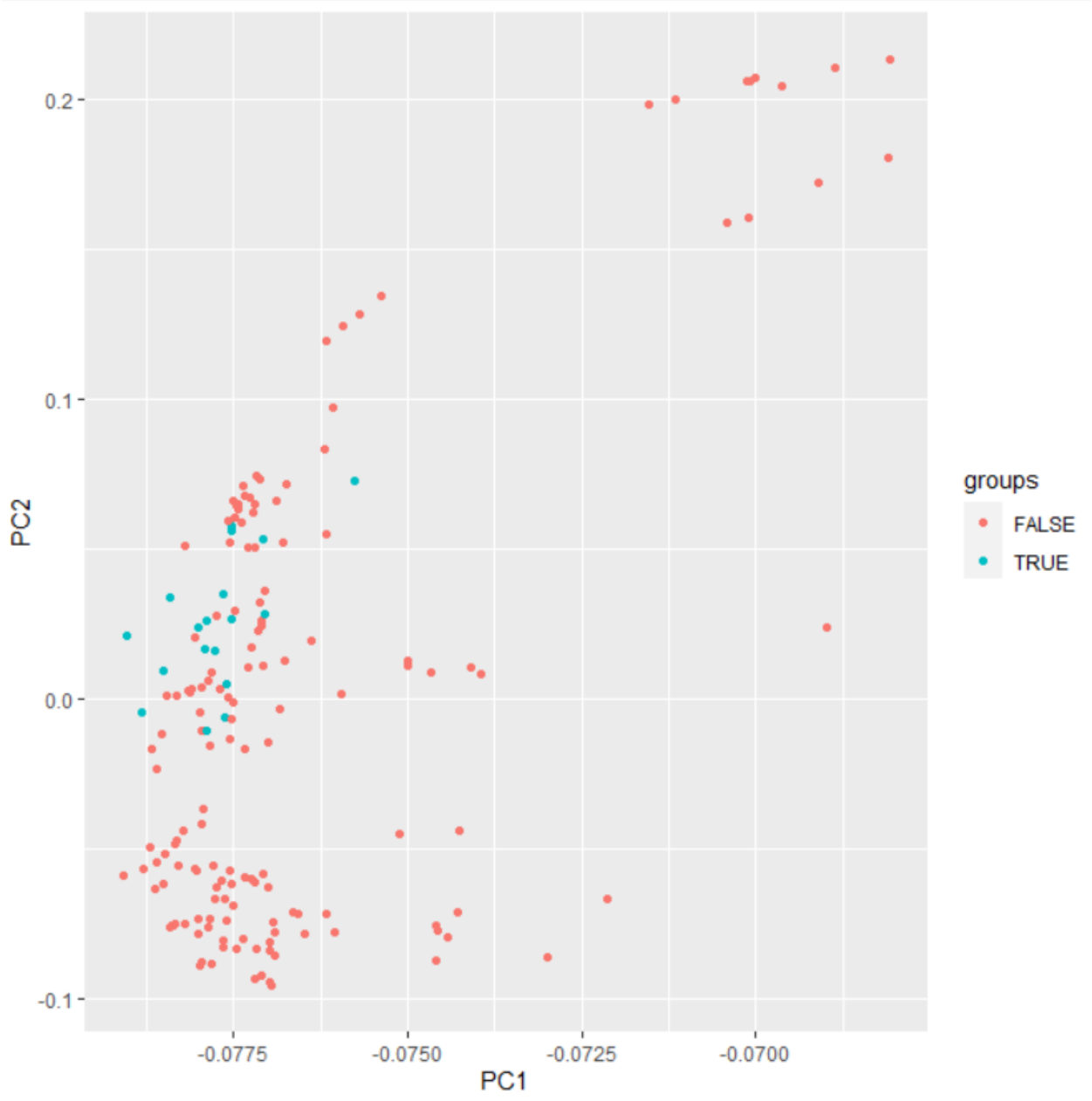
حال همین کار را برای pca و tsne استفاده میکنیم.

## 2.3 PCA

نمودار خروجی PCA با درنظر گرفتن تمامی source name ها به صورت زیر میباشد :

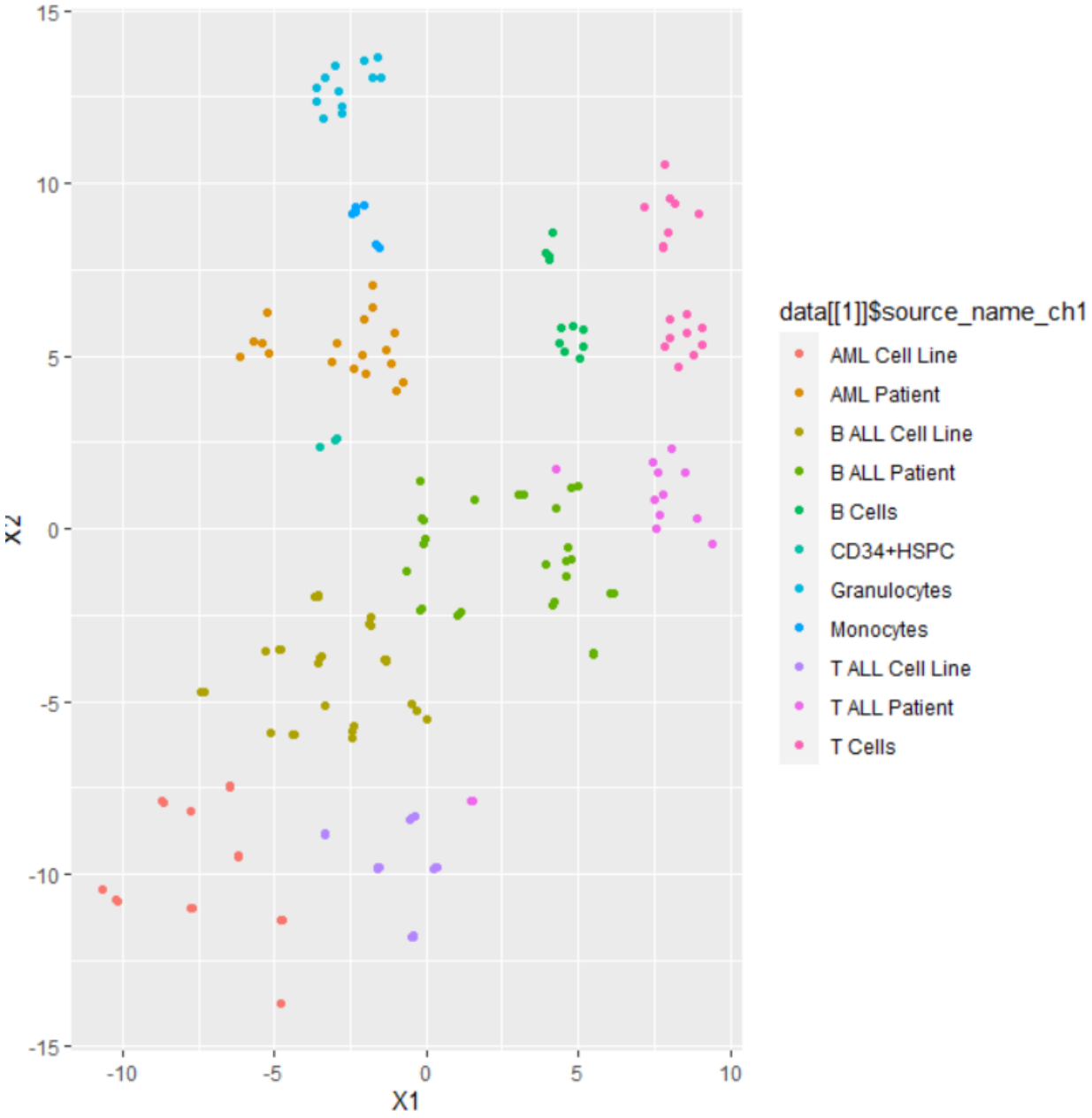


همانطور که ملاحظه میشود این متود توانایی جدا کردن داده ها به خوبی UMAP را نداشته است. اگر بخواهیم به صورت گروه های سالم و ناسالم به داده‌ها نگاه کنیم به نمودار زیر میرسیم :

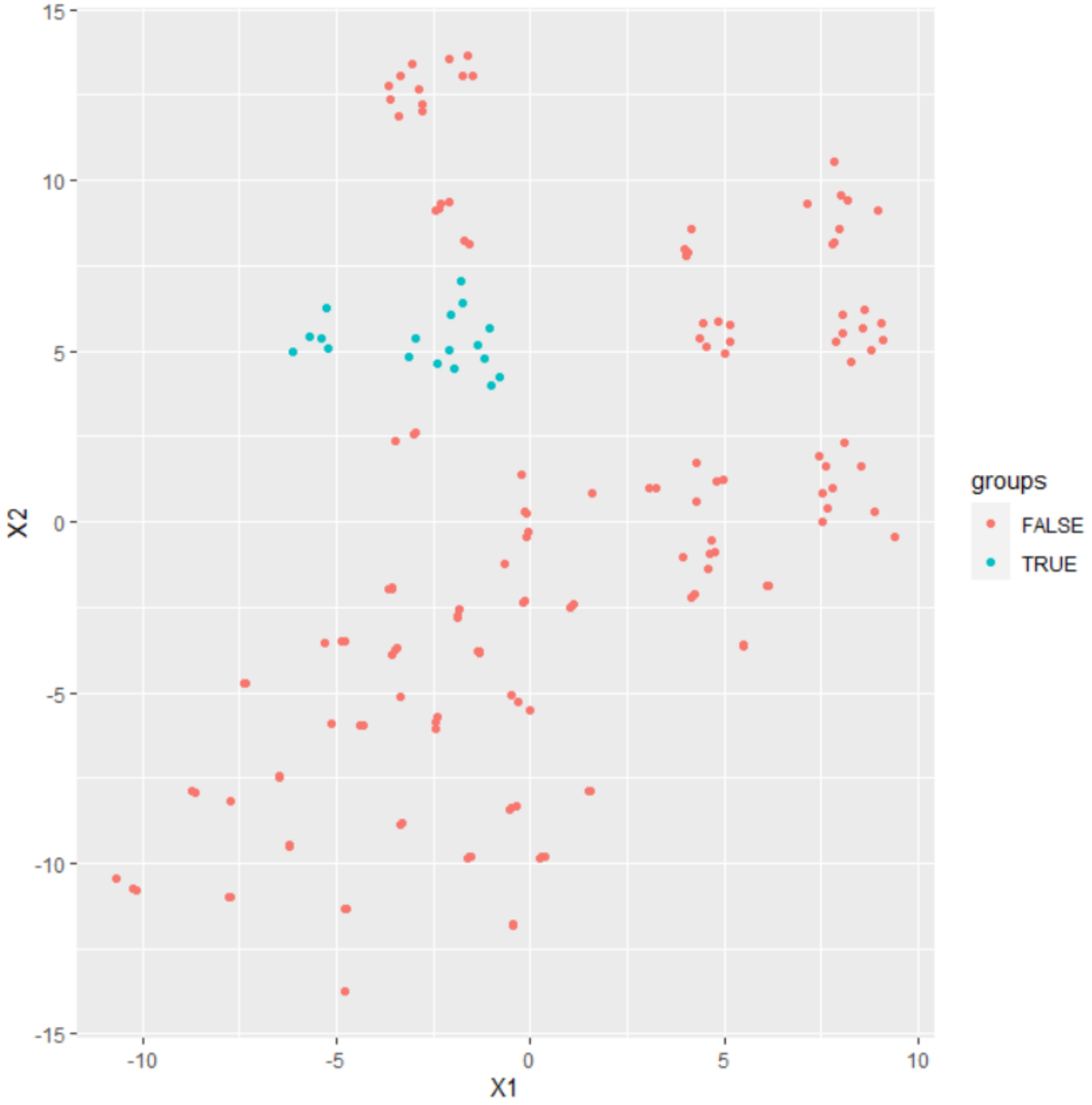


## 3.3 TSNE

نتایج روش TSNE بر روی داده‌ها به صورت زیر است :



در این نمودارها رنگ داده‌ها بر اساس source name آنها مشخص شده است. نموداری که در آن رنگ داده‌ها بر اساس سالم یا ناسالم بودن آنها باشد به صورت زیر میباشد.

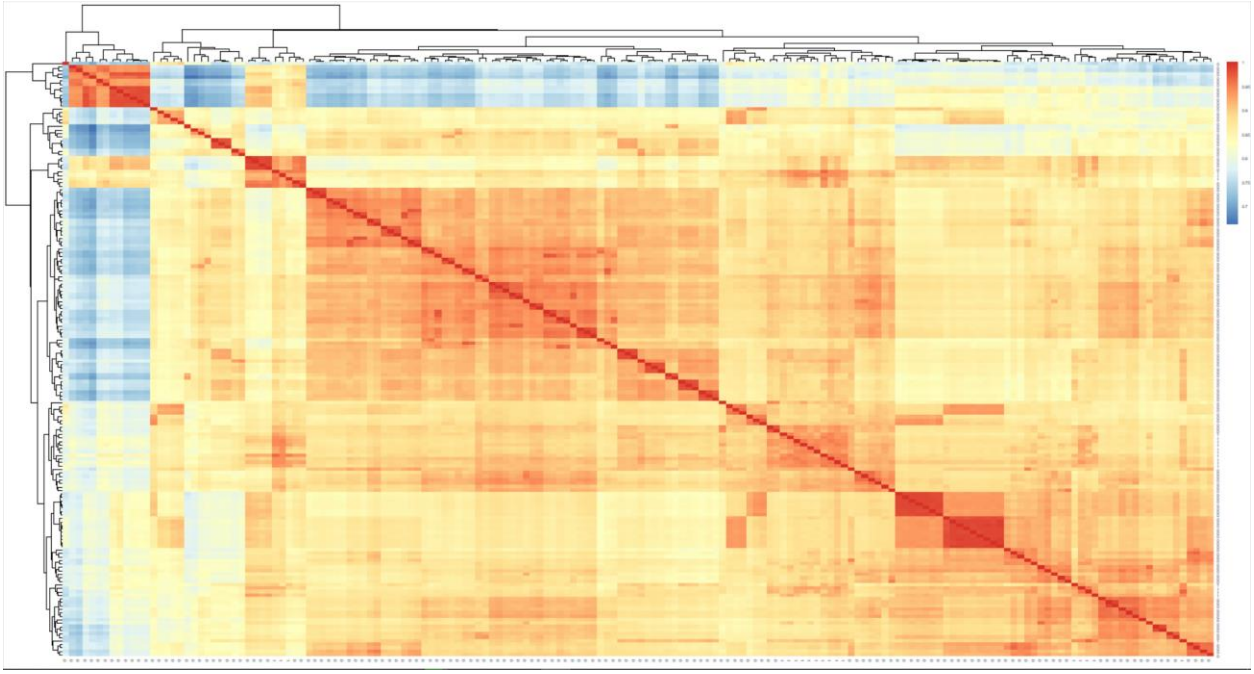



همانطور که مشاهده میشود نقاط از هم تفکیک پذیر هستند.

در بین این ۳ روش، UMAP و TSNE توانسته‌اند داده‌ها را به خوبی گروه بندی کنند بنابراین بنظر میرسد که این دو روش برای اجرای کاهش ابعاد بر روی این داده‌ها مناسب بوده اند.

## ۴ تحلیل نمونه‌ها

ستون source name نوع نمونه‌ای است که در اختیار داریم. مقدار AML patient برای نمونه‌های ناسالم و بقیه نمونه‌ها مربوط به انواع مختلف نمونه‌های سالم هستند. در اینجا میخواهیم کوریلیشن بین آنها را محاسبه کنیم. نمودار زیر نمودار هیت مپ correlation بین داده‌ها است. همانطور که ملاحظه میشود مقادیر روی قطر بیشترین مقدار ( 1 ) را دارند و قسمت هایی که به رنگ آبی هستند کوریلیشن کمتری وجود دارد.



این نمودار به طور دقیقتر در فایل cors.png قرار دارد. همانطور که در این فایل و نمودار UMAP مشاهده شد، داده‌های سالمی وجود دارند که به داده‌های ناسالم در 2 بعد نزدیکتر هستند و همچنین در این نمودار هم کوریلیشن بیشتر با آنها دارند. بنابراین به نظر میرسد که در مراحل بعدی تست میتوانیم از این نمونه‌های سالم تست های بیشتری بگیریم.