

**Amirkabir University of Technology  
(Tehran Polytechnic)**



**Department of  
Computer Engineering**

# **Course : Statistical Pattern Recognition**

## **Homework 4**

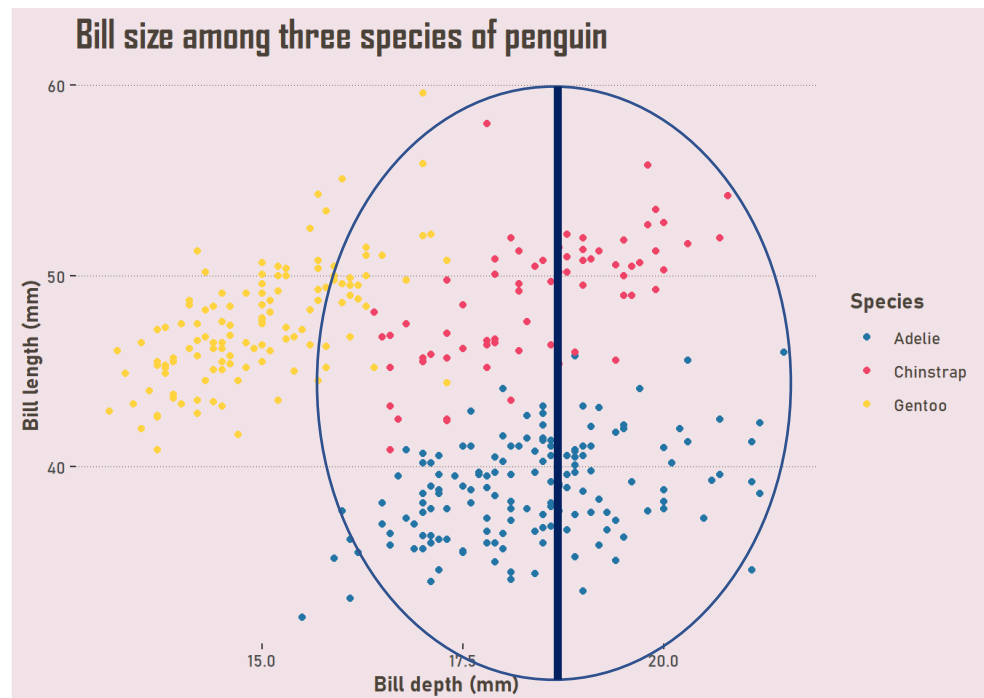
**Najmeh Mohammadbagheri**

**99131009**

## گزارش تمرین

### سوال یک

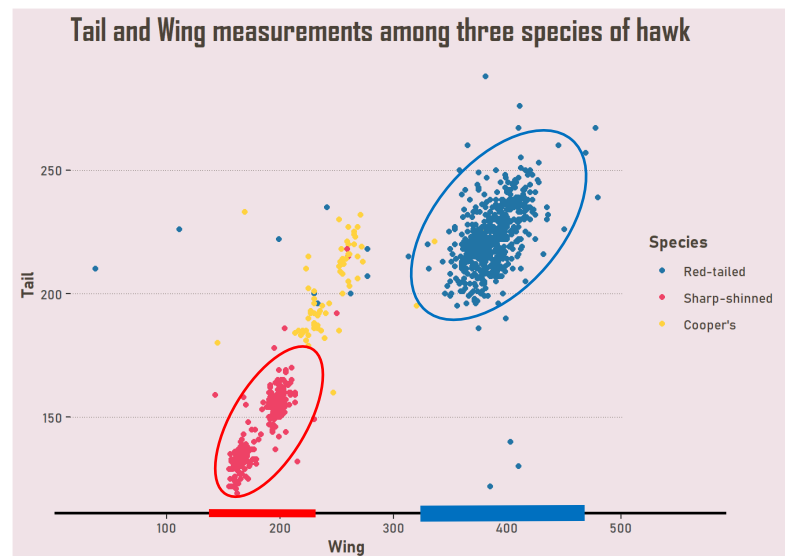
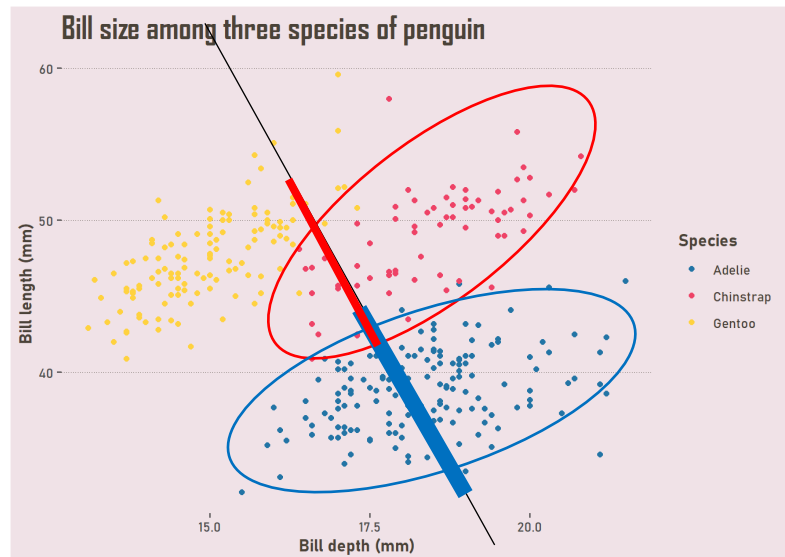
قسمت (a) توزیع داده‌ها دوکلاس آبی و قرمز با بیضی آبی رنگ مشخص شده است. قطر بزرگ این بیضی، خطی که داده‌ها روی آن تصویر میشوند را نشان می‌دهد. (با روش **pca** بیشترین واریانس در نظر گرفته میشود که این خط آبی هم همان بیشترین جهت واریانس است.



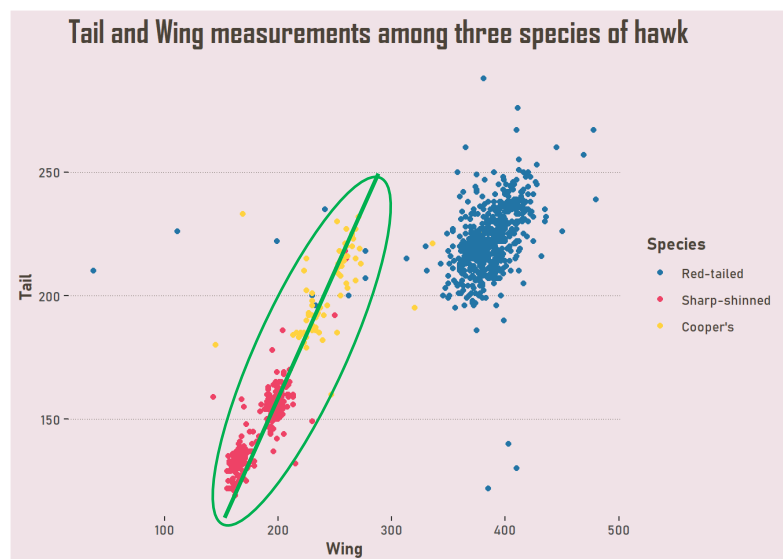
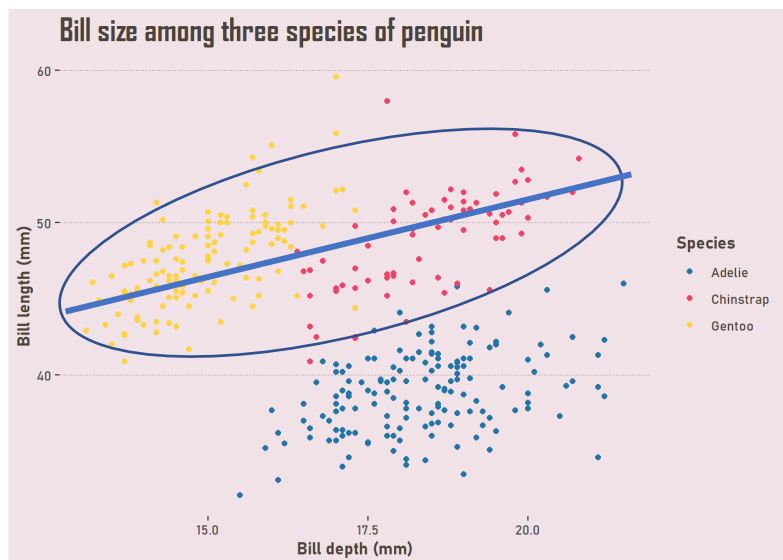
در تصویر دوم توزیع داده‌ها دوکلاس آبی و قرمز با بیضی سبز رنگ مشخص شده است. قطر بزرگ این بیضی، خطی که داده‌ها روی آن تصویر میشوند را نشان می‌دهد. (با روش **pca** بیشترین واریانس در نظر گرفته میشود که این خط سبز هم همان بیشترین جهت واریانس است.



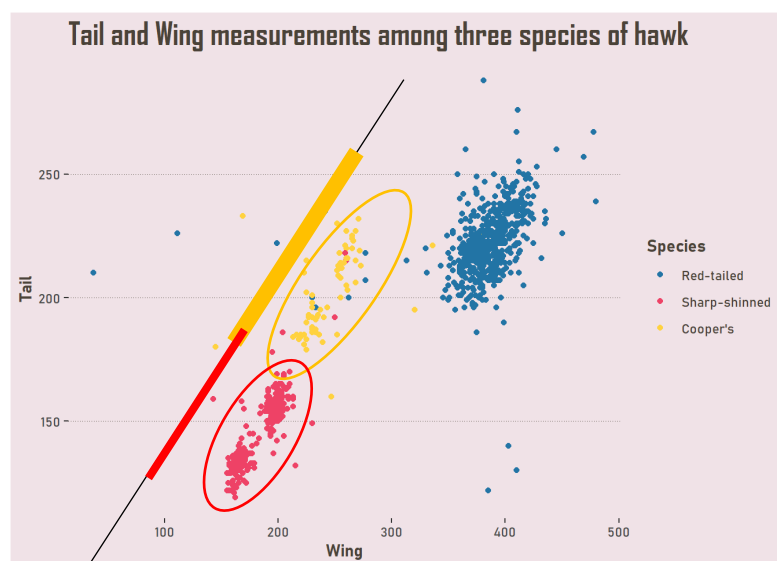
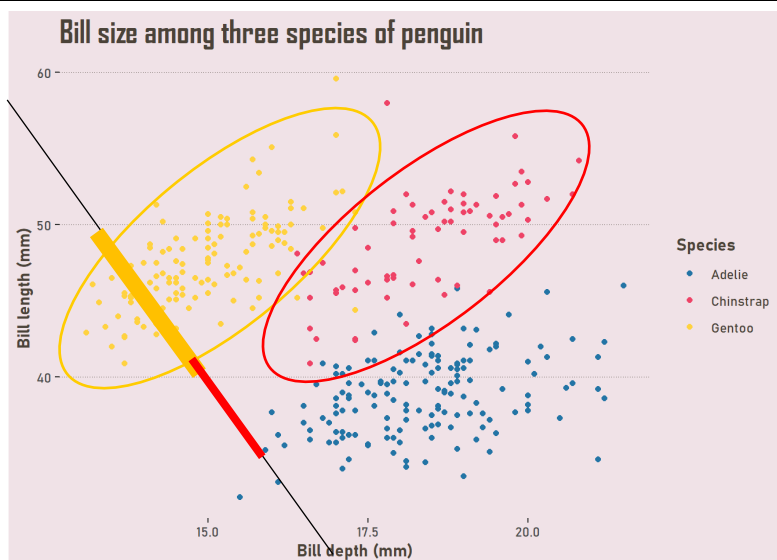
قسمت (b) توزیع هر کلاس با رنگ متناظرش مشخص شده است. خط مشکی جهتی را نشان میدهد داده ها با روش فیشر بر آن تصویر میشوند. این خط بهترین جهت است چون میانگین توزیع ها در فضای جدید بیشتری فاصله را دارند و کمترین واریانس.



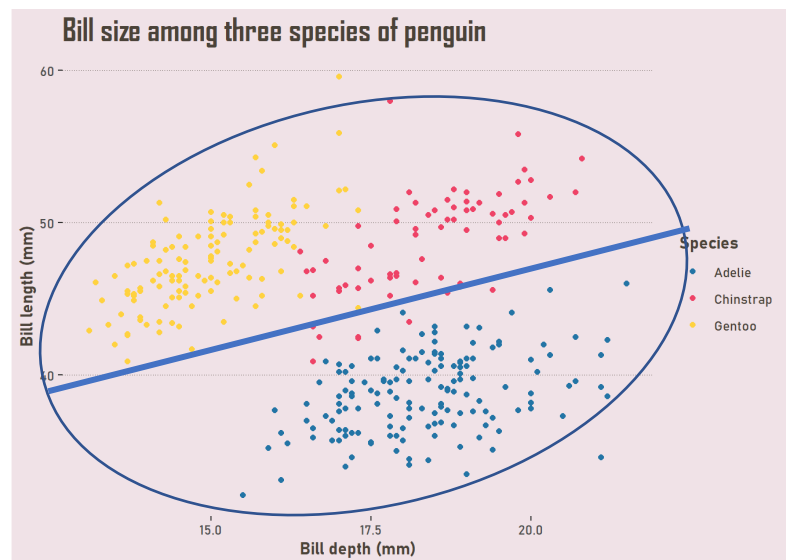
قسمت (C)



قسمت d)

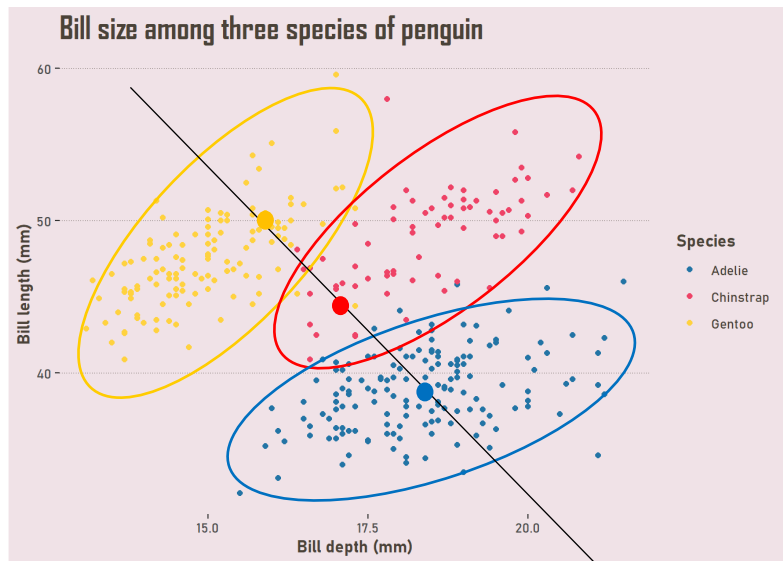


قسمت (e)

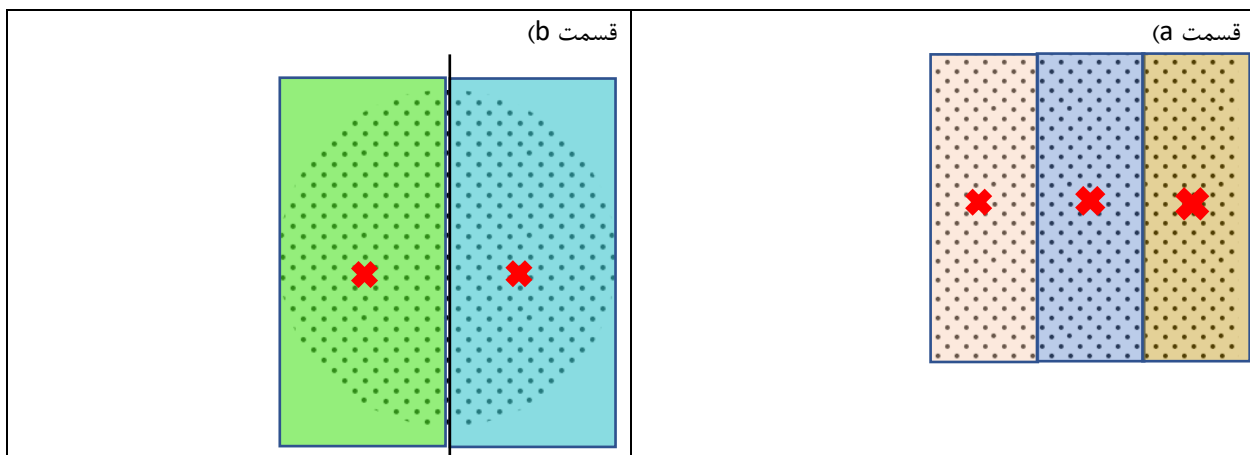


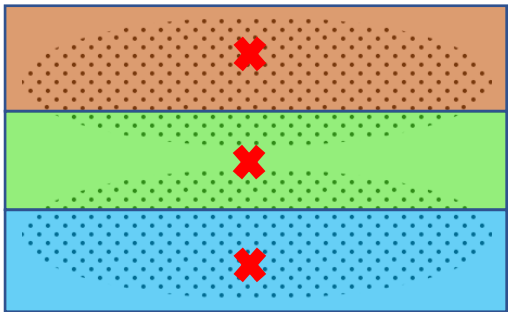
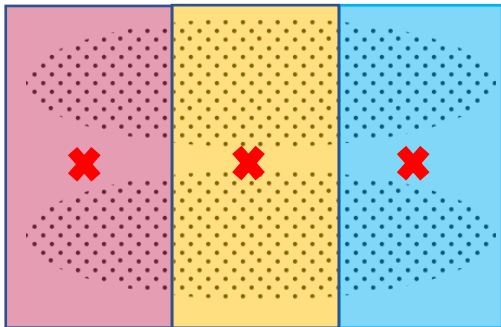
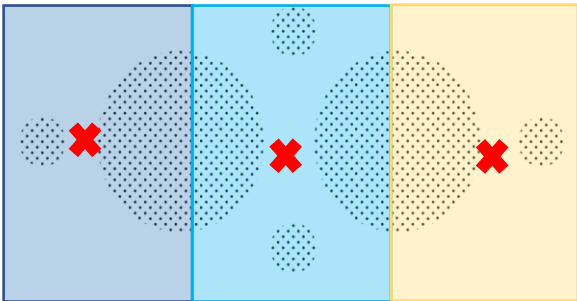
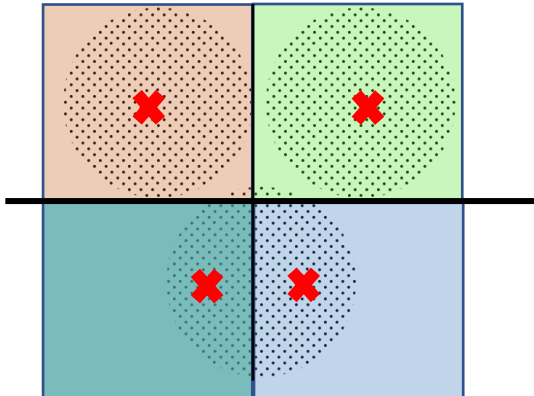
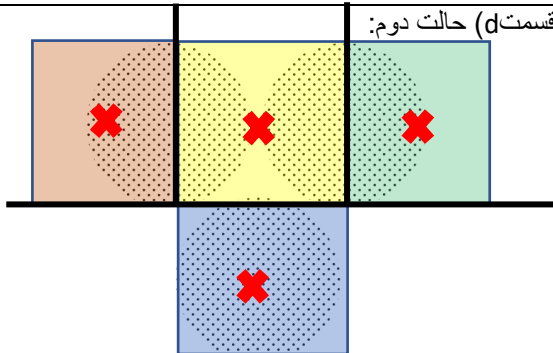
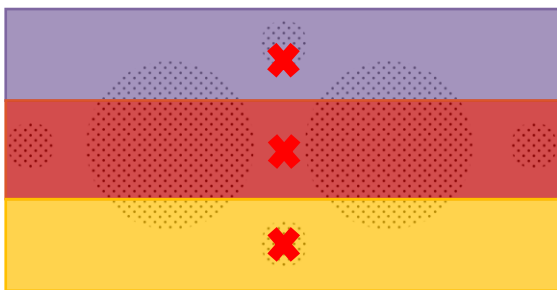
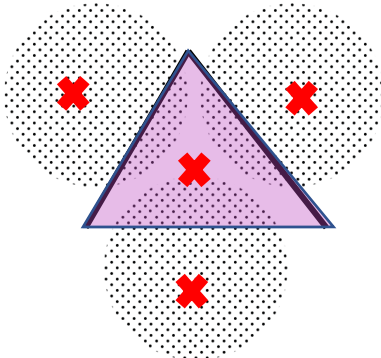
قسمت f)

توزیع هر کلاس با رنگ متناظرش مشخص شده است. خط مشکی جهتی را نشان میدهد داده ها با روش فیشر بر آن تصویر میشوند. این خط بهترین جهت است چون میانگین توزیع ها در فضای جدید بیشتری فاصله را دارند و کمترین واریانس. (اگر میخواستیم واریانس را نشان دهیم خیلی درهم رفتگی رنگ ها پیش می آمد، به همین دلیل صرف نظر کردم).



## سوال دوم



<p>قسمت C) حالت دوم: حالت بهینه</p> 	<p>قسمت C) حالت اول:</p> 
<p>قسمت e) حالت اول:</p> 	<p>قسمت d) حالت اول: حالت بهینه</p> 
<p>قسمت d) حالت دوم:</p> 	<p>قسمت e) حالت دوم:</p> 
	<p>قسمت d) حالت سوم:</p> 



(f)

در شکل راست که داده‌های در و پنجره را داریم. در الگوریتم‌های خوشه‌بندی ای که تاکنون بحث کرده ایم فرض بر این است که داده‌ها را داریم و با حرکت بر روی داده‌ها خوشه‌ها را پیدا می‌کنیم. نه با پیدا کردن ناحیه‌ای هایی که داده نیست.

(g)

محدودیت‌ها:

شکل توزیع داده‌ها بسیار مهم است در تصمیم گیری برای مدل فاصله. همچنین حد آستانه‌ای که در نظر گرفته میشود که در نهایت خوشه‌ها بر اساس آن جدا شوند.

## سوال سوم

قسمت (a)

از فرمول زیر برای محاسبه‌ی خطای نمایش استفاده شد:

$$\sum_{j=1}^n \|x_j\|^2 - \sum_{i=1}^k e_i^t S e_i$$

و خطای حاصل ۱۹.۳۶ شد که بنظر میزان خوبی است. پس میتوان گفت سه بردار ویژه‌ی اول برای نمایش داده‌ها خوب هستند.

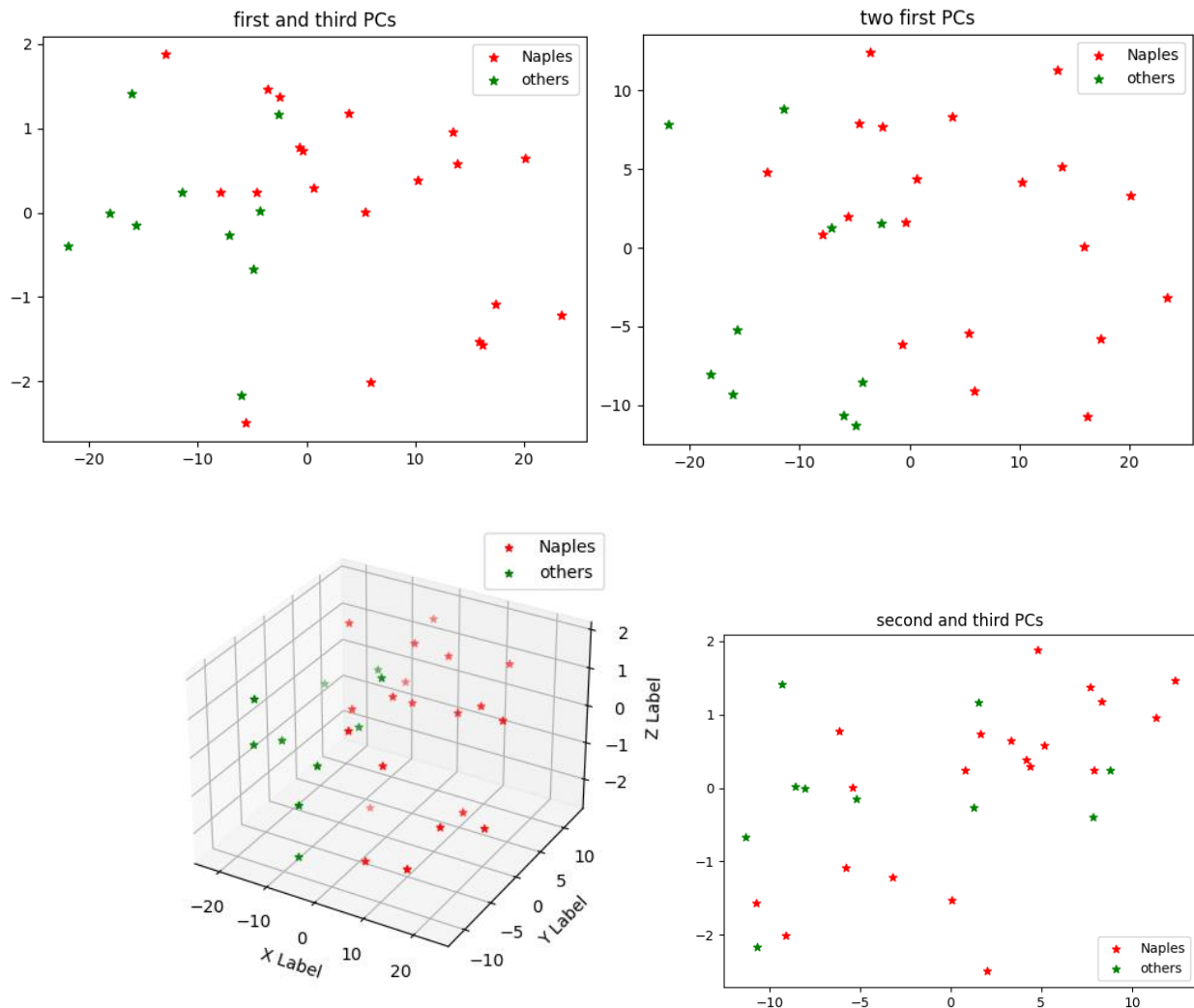
```
representation error:
19.36137114579833
```

دو مقدار ویژه‌ی بزرگتر برابر با ۴۲۲۲ و ۱۵۴۲ (گرد شده اند اعداد) و سومین بزرگترین مقدار ویژه ۲۶ است. سه تای بعدی ۸ و ۱ و ۰.۳ هستند که میبینیم مقادیر خیلی کوچکی هستند، پس اگر از آنها استفاده نشود خطا خیلی کم میشود.

البته میتوانستیم هم تنها از دو مقدار ویژه‌ی اول استفاده کنیم که در این حالت خطای نمایش ۵۷.۹ میشود. (چون فاصله‌ی ۲۶ تا ۱۵۴۲ خیلی زیاد است گفتم میشود از ۲۶ استفاده نکرد).

```
representation error:
57.925011105784506
```

قسمت (b) امتیازی



همانطور که از تصویر بالا مشخص است داده ها تفکیک پذیر نیستند. این مطلب با قسمت اول مغایرتی ندارد. در قسمت اول گفتیم خطای نمایش خوب است. یعنی نمایش داده ها در بعد پایین خوب است، نه اینکه داده ها از لحاظ کلاس خوب جدا شده اند. اگر میخواستیم بعد داده ها را برای مسالهی کلاس بندی کاهش دهیم باید از فیشر استفاده می کردیم. PCA تنها برای نمایش است.

## سوال چهارم

به اولین سطر داده ها یک خط اول اضافه شد که سطر اول داده ها کم خوانده نشود.

(a) تعداد ۲۰۰ تا کامپوننت اول ماتریس Y را برای PCA در نظر گرفتیم و مقادیر ویژه را مقایسه کردم. دو کامپوننت اول اختلاف بیشتری با بقیه ی مقادیر ویژه داشتند. به همین دلیل دو کامپوننت اول برای مسالهی PCA انتخاب شد.

[13.50105143 11.00079584 8.8615011 7.8588508 6.40080316 6.04214501 5.69805717 5.60152651  
5.46710822 5.4308534 5.40247067 5.32393163 5.29818259 5.29512084 5.25837858 5.23751293

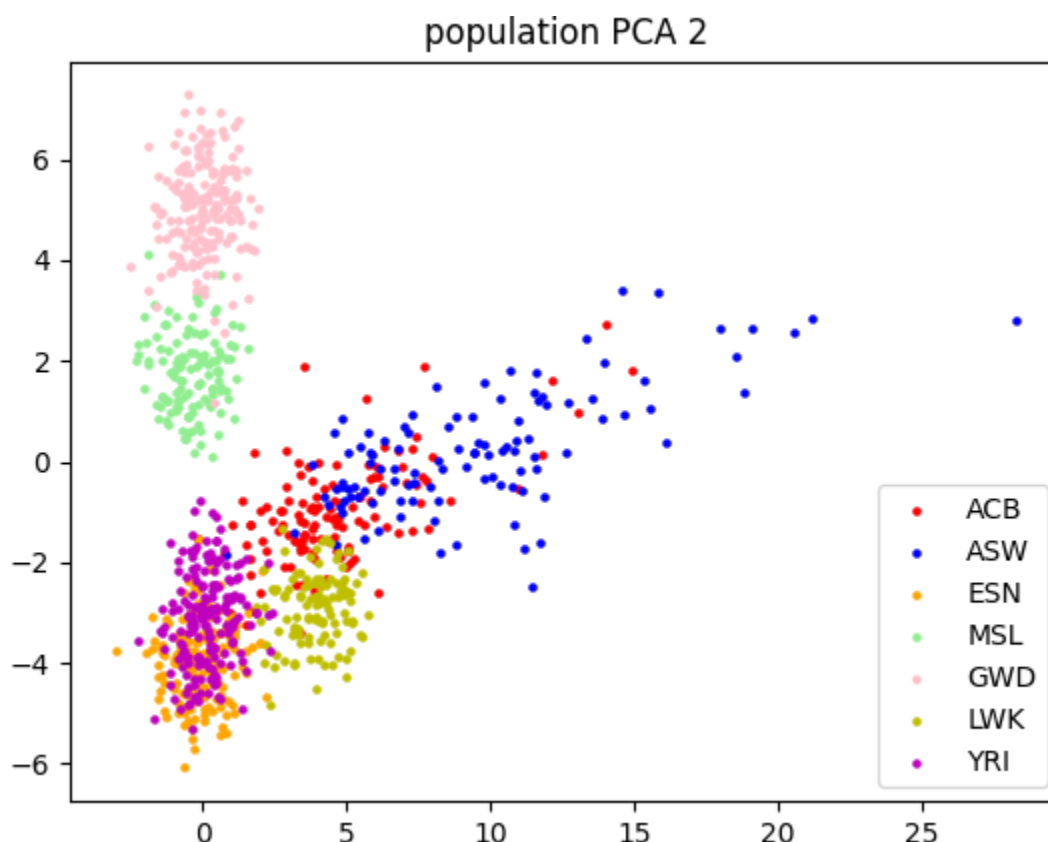
5.18546432 5.17461348 5.15575759 5.12911983 5.11599402 5.09864761 5.07269172 5.05933126  
5.03468856 5.02556544 4.99878677 4.9694915 4.95208917 4.94540348 4.93745863 4.91621812  
4.898328 4.8946895 4.87881365 4.86743119 4.85173371 4.83917197 4.8269187 4.81420909  
4.80721851 4.78333838 4.77188329 4.76436563 4.75905997 4.74034193 4.72337743 4.71379688  
4.69840382 4.69479501 4.68263826 4.65675147 4.64485999 4.63909219 4.62348299 4.61532894  
4.5980376 4.58327348 4.57663415 4.56189688 4.55314724 4.54553388 4.53554325 4.50909885  
4.50618566 4.4930243 4.48552797 4.47184806 4.46686867 4.4553432 4.45197489 4.44322025  
  
4.41730148 4.41128552 4.39097279 4.38047243 4.37574591 4.37306548 4.35755729 4.34963049  
4.34001386 4.33488423 4.32873734 4.3193468]

بنابراین بُعد ماتریس  $Y$  بعد از تبدیل، ۹۹۵ در ۲ میشود.

(b)

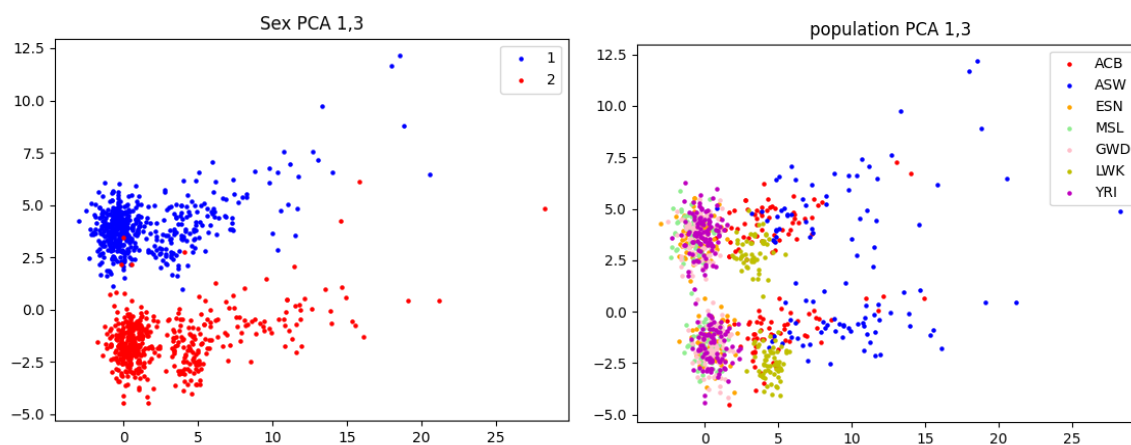
این داده‌ها برای ۷ جمعیت است: ACB, ASW, ESN, MSC, GWD, LWK, YRI

دیگر جمعیت‌ها در دیتاست وجود نداشت.



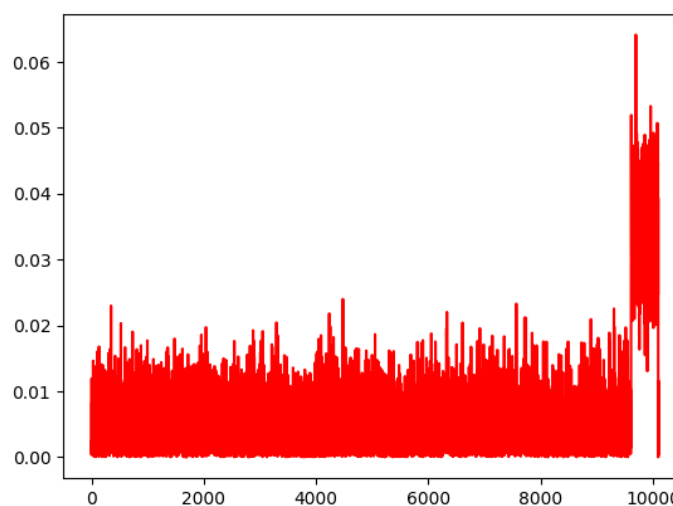
(c) اگر این مناطقی که در نمودار قسمت قبل مشخص شده اند را بر روی نقشه نگاه کنیم قسمت هایی که از لحاظ جغرافیایی بهم نزدیک تر هستند در خوشه بندی هم بهم نزدیک تر هستند یعنی خوشه‌هایشان بهم نزدیک تر است. دلیل این امر این است که شرایط محیطی این مناطق بیشتر بهم نزدیک بوده و این بر ژنتیک نیز تاثیر دارد.

(d,e)



کامپوننت سوم برای جنسیت است، زیرا با در نظر گرفتن لیبل گذاری با متادیتا جنسیت، تفکیک پذیری بسیار خوبی از داده‌ها گرفتیم. داده‌های قرمز مربوط به زن و داده‌های آبی مربوط به مرد است. اما اگر شکل اول را در نظر بگیریم میبینیم تقسیم بندی بر اساس موقعیت جغرافیایی در این حالت خوشه‌های مجزا و خوبی را به ما نمی‌دهد.

(f)



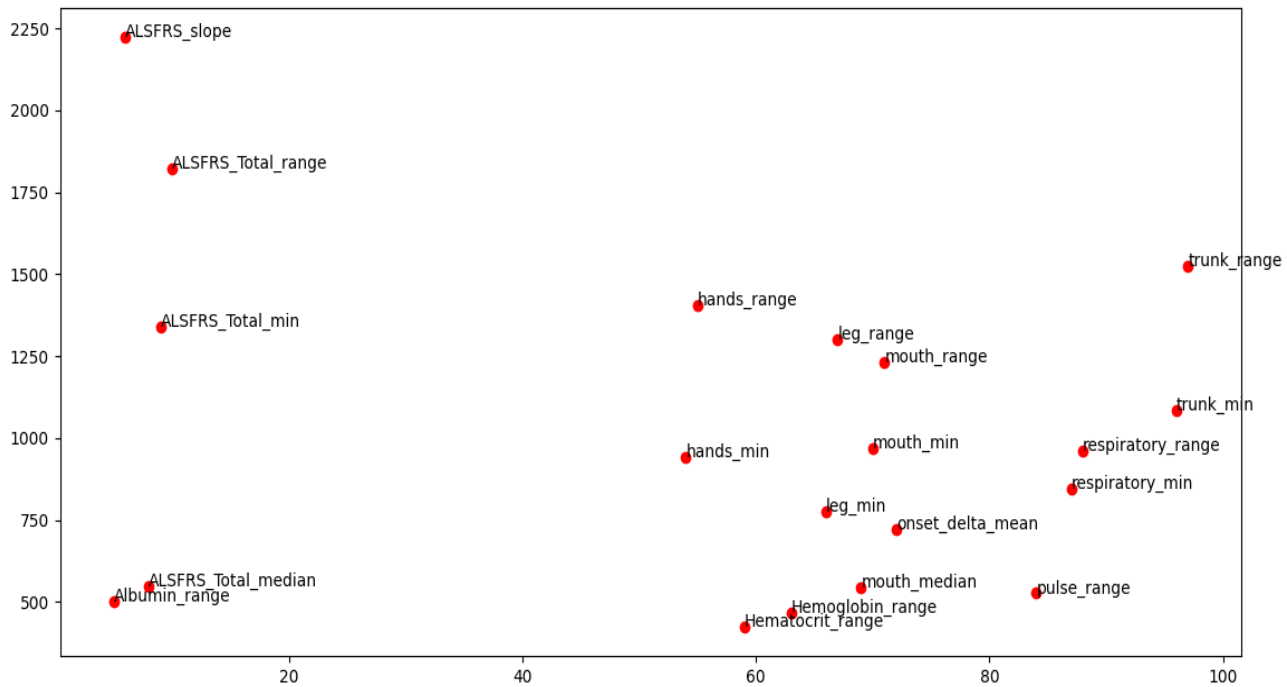
در قسمت قبل دیدیم که کامپوننت سوم و اول جنسیت را خوب تعیین میکردند و در این قسمت میبینیم که کامپوننت سوم بدین صورت است. بنابراین نتیجه می‌گیریم این ویژگی برای تعیین جنسیت است و قسمت آخر نوکلتوبیس تاثیر اصلی را بر جنسیت دارد.

## سوال پنجم

- (a) مراکز نهایی در پوشه تمرین قرار دارد. تکرار حلقه دوبار بود.
- (b) بهترین الگوریتم K means++ است. در این الگوریتم هر مرحله یک مرکز انتخاب می شود و در انتخاب مرکز بعدی سعی میشود بیشترین فاصله را از مراکز فعلی داشته باشد. این الگوریتم پیاده سازی شده است. مراکز اولیه بدست آمده با این روش با اسم centroids\_b.txt ذخیره شده اند. تکرار حلقه در این حالت نیز دوبار بود.

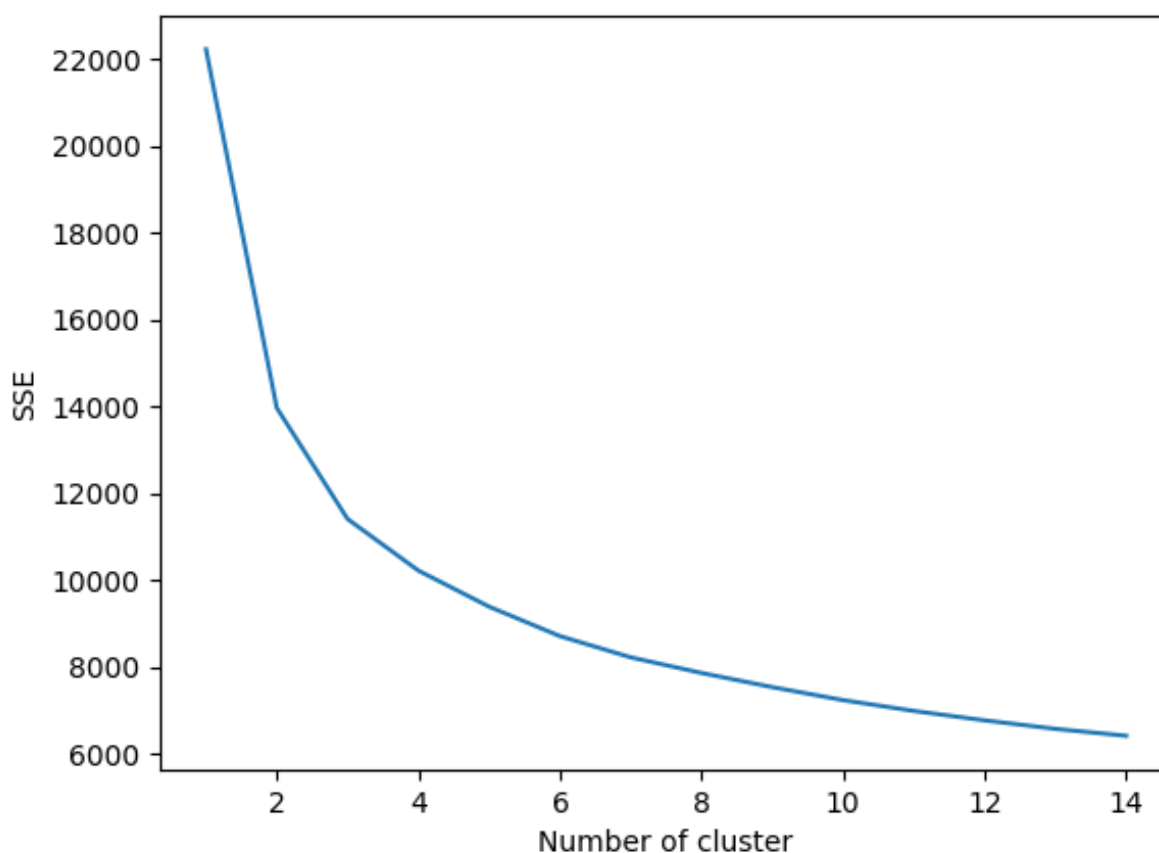
## سوال ششم

- (a) ابتدا نرمال سازی داده ها انجام شد و سپس به ازای تمام ویژگی ها کواریانس با ویژگی ALSFRS\_slope محاسبه شد. در آخر نیز ۱۰ تا کواریانس بزرگتر انتخاب شد.



```
ALSFRS_slope 2222.999999999999
ALSFRS_Total_range 1821.3144340885356
trunk_range 1525.8834300311057
hands_range 1405.7150303168319
ALSFRS_Total_min 1337.823284634067
leg_range 1299.3597432955967
mouth_range 1229.7508435270208
trunk_min 1086.0024822859843
mouth_min 968.2011400809444
respiratory_range 962.2886674966344
hands_min 942.7905403903893
```

(b)



بهترین  $k$  برابر با ۳ است. زیرا بعد از آن تغییرات SSE به شیب کمی همراه است.

(C)

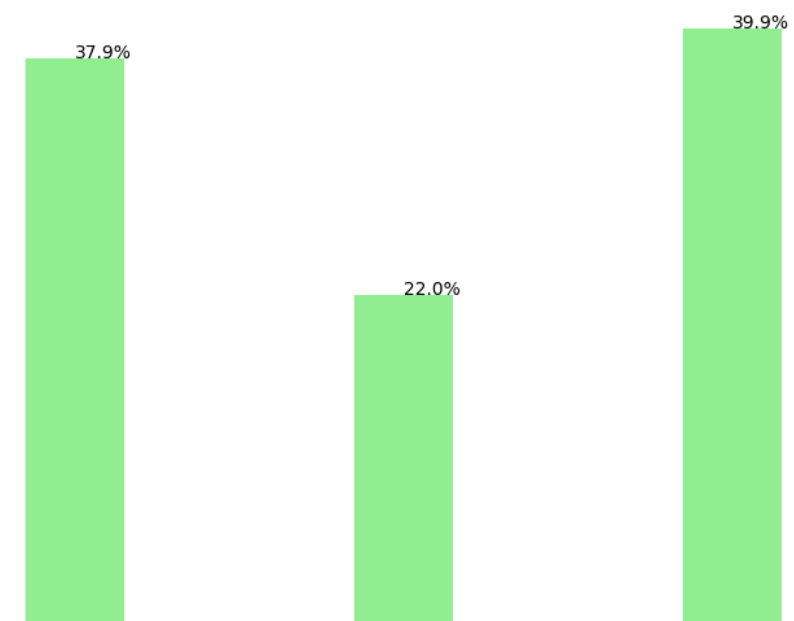
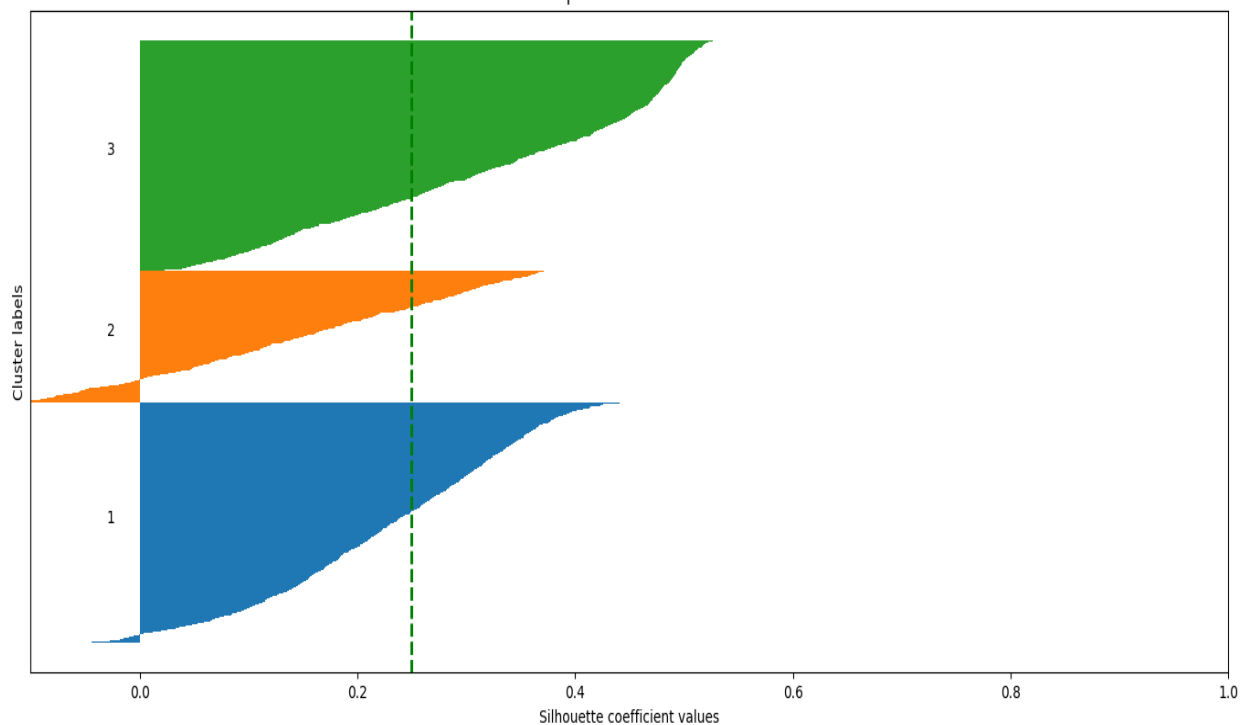
ضریب silhouette میزان کیفیت خوشه بندی را اندازه گیری می کند و مطابق شکل زیر محاسبه می شود.

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}}$$

$a$  بیانگر میانگین فاصله ی نقطه  $o$  تا سایر داده های هم خوشه ای و  $b$  بیانگر میانگین فاصله ی نقطه  $o$  تا داده های خوشه های دیگر است. در نمودار زیر این ضریب به ازای تمامی نقاط محاسبه شده است. نقاط در محور عمودی و ضرایب در محور افقی قابل مشاهده اند. خط نقطه چین میانگین تمام این ضرایب است. هر چه این خط به یک نزدیک تر باشد بهتر است. یک بودن یعنی داده های داخل خوشه ها بسیار بهم نزدیک هستند و خوشه ها از هم بسیار دور هستند.

### Silhouette analysis using $k = 3$

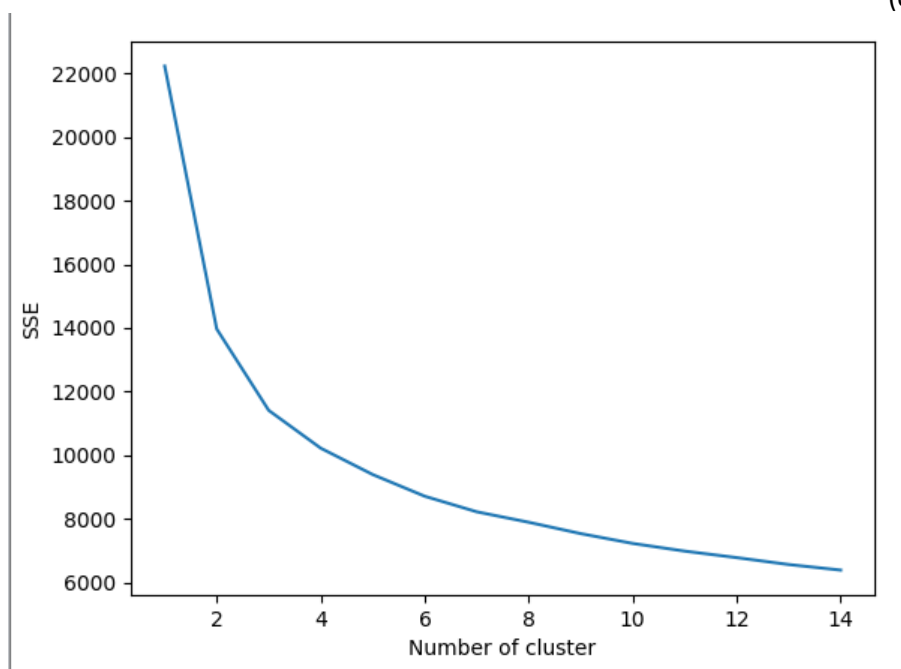
Silhouette plot for the various clusters



چون شماره‌ی خوشه‌ها اهمیت ندارد، شماره‌ها مشخص نشده‌اند.



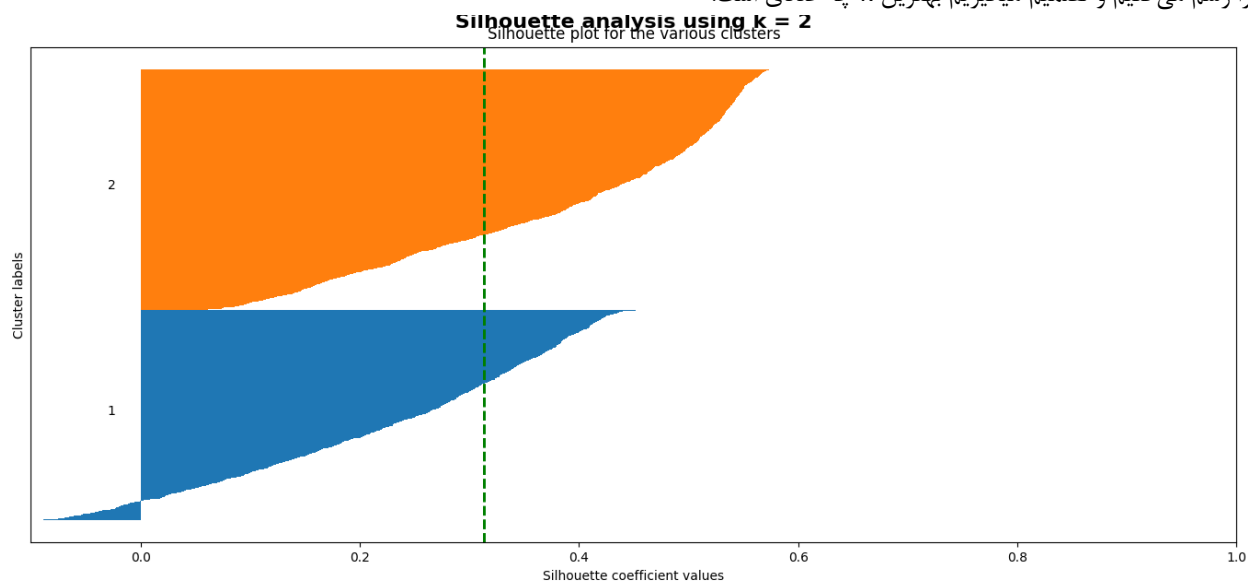
(d)

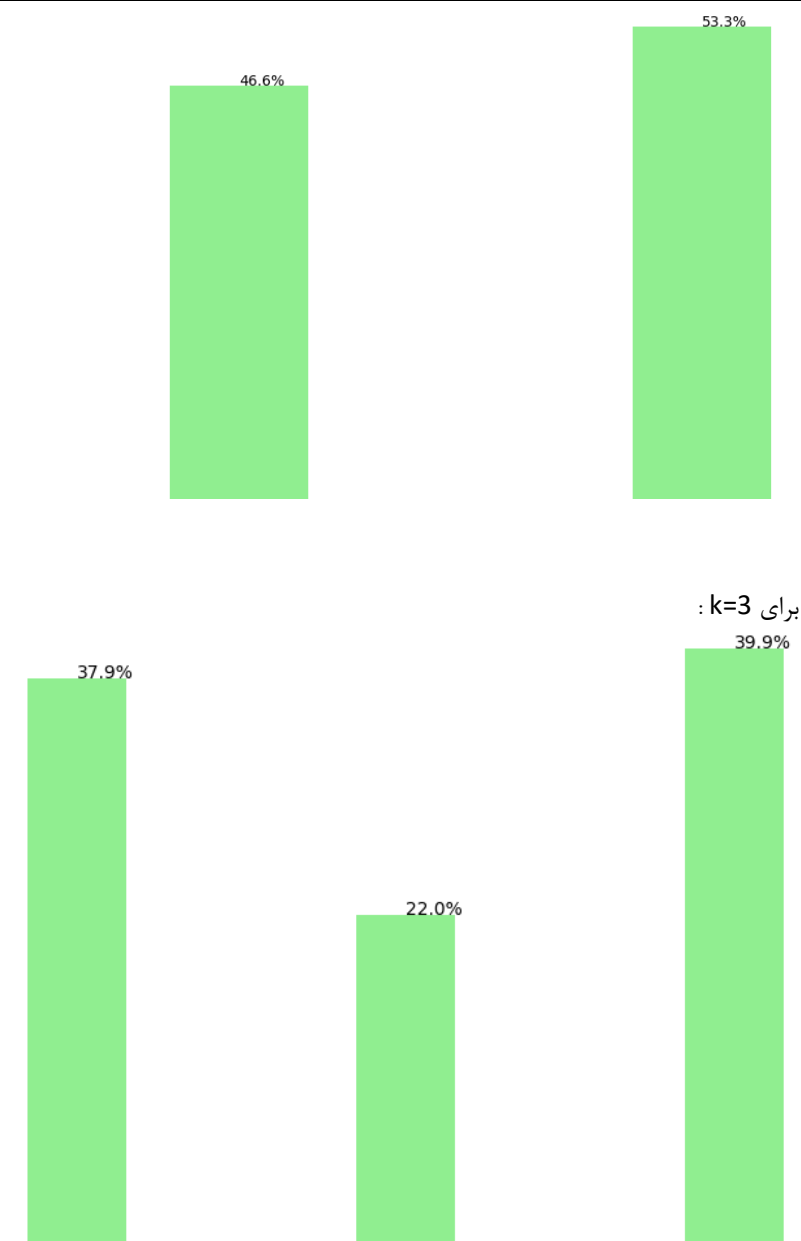


همانند قسمت b بهترین مقدار پارامتر ۳ است.

(e)

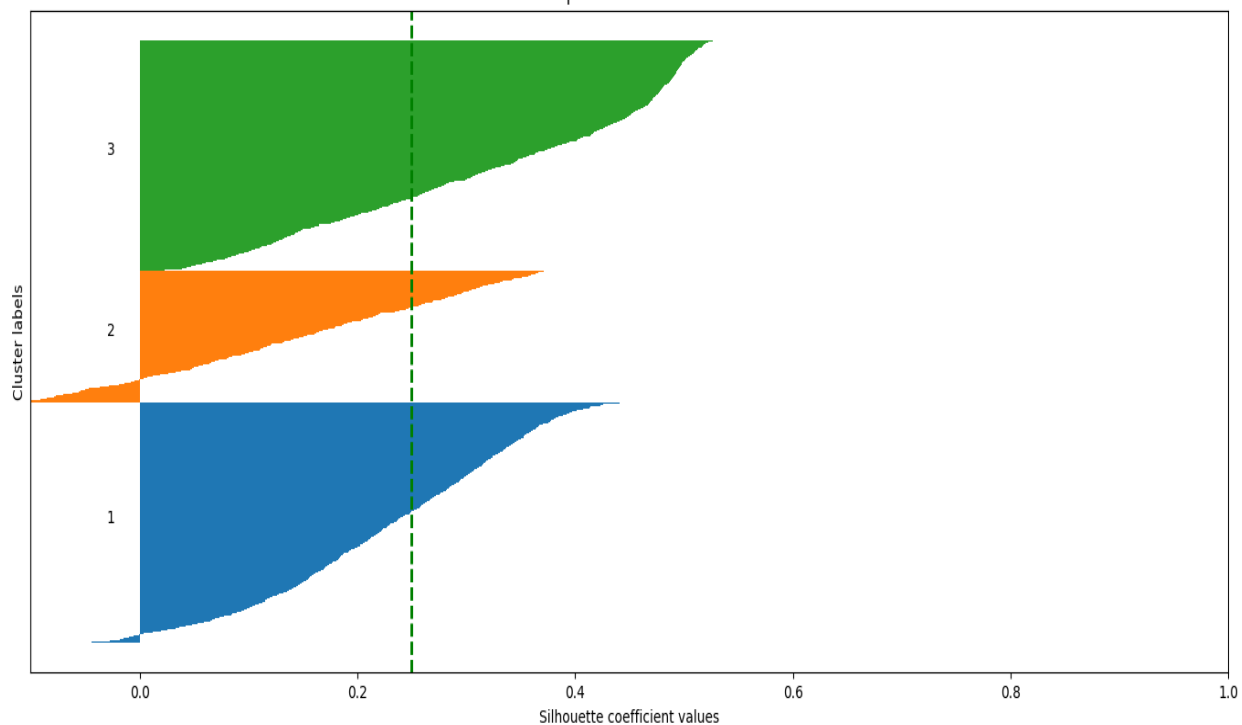
با تعاریفی که در قسمت c کردیم و دانشی که نسبت به نمودار میله ای داریم (بیانگر خلوص هر خوشه است) برای  $k = 2, 3, 4, 5$  این نمودارها را رسم می کنیم و تصمیم میگیریم بهترین  $k$  چه عددی است.





### Silhouette analysis using $k = 3$

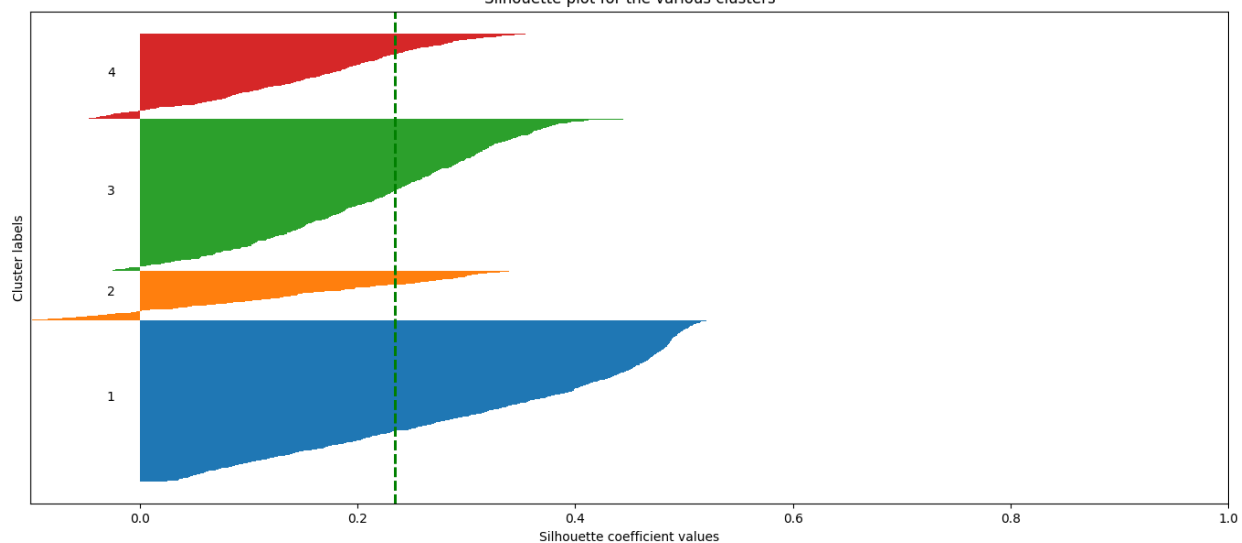
Silhouette plot for the various clusters

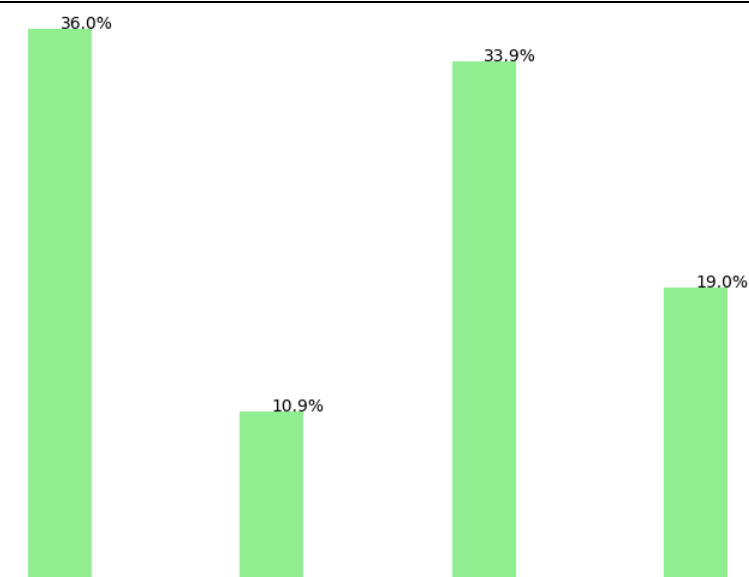


اکنون برای  $k=4$ :

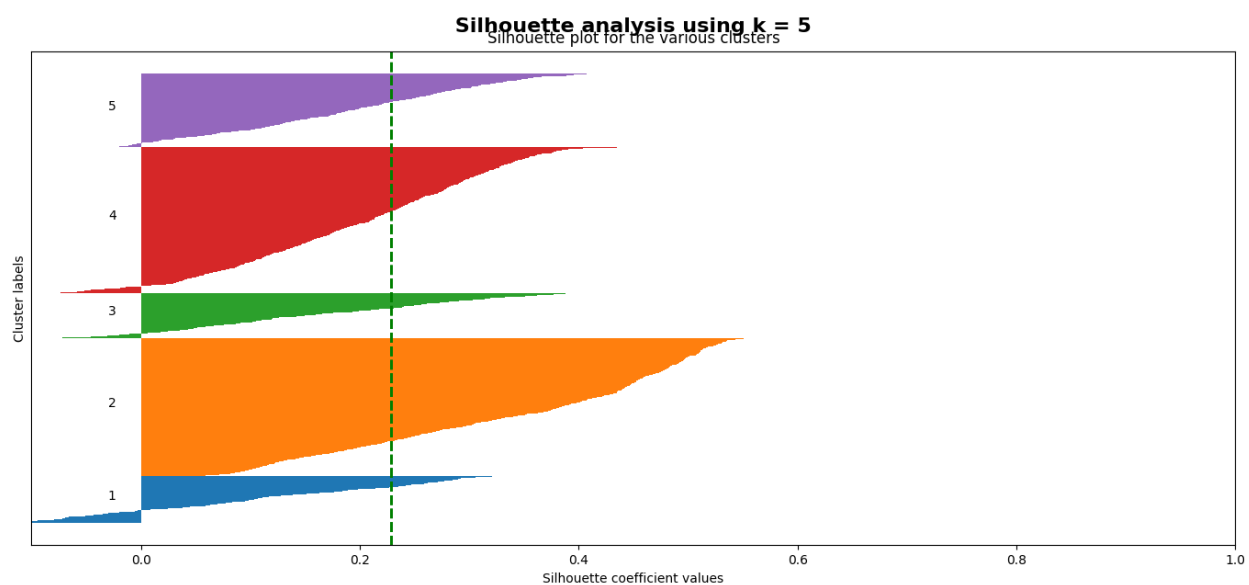
### Silhouette analysis using $k = 4$

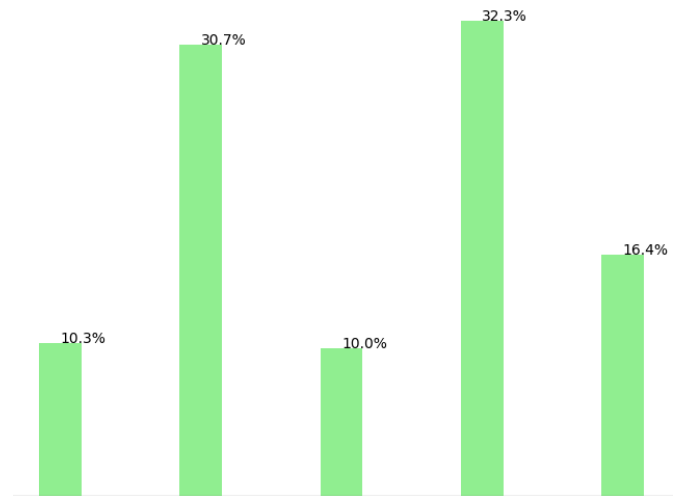
Silhouette plot for the various clusters





اکنون برای  $k = 5$ :

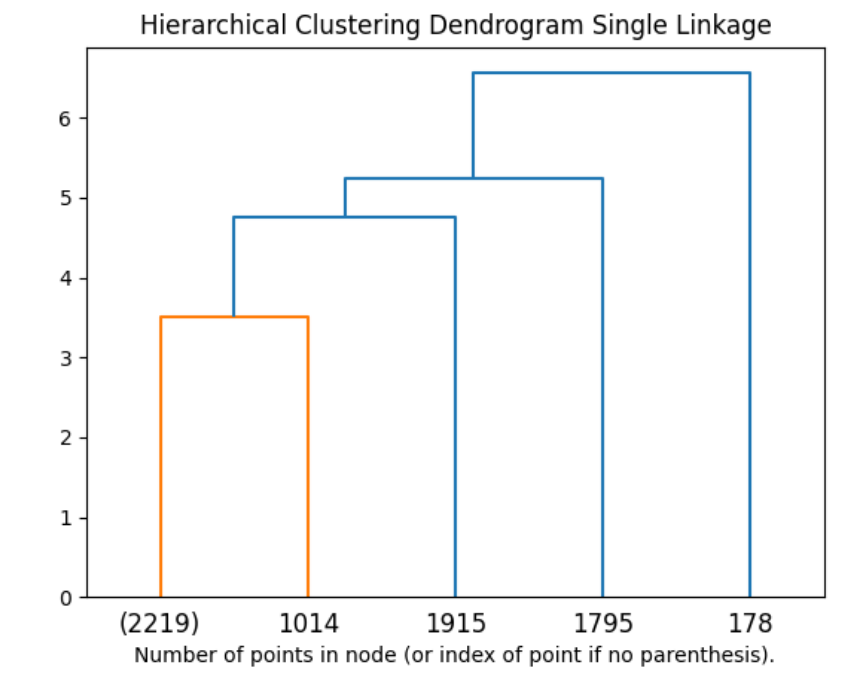


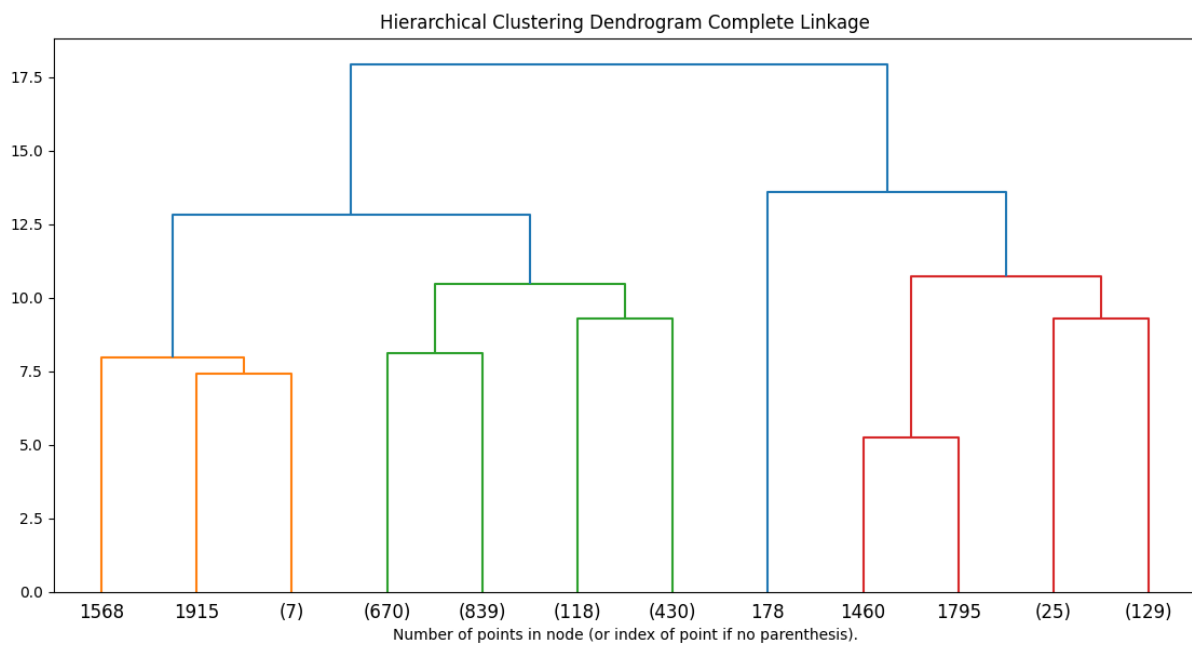
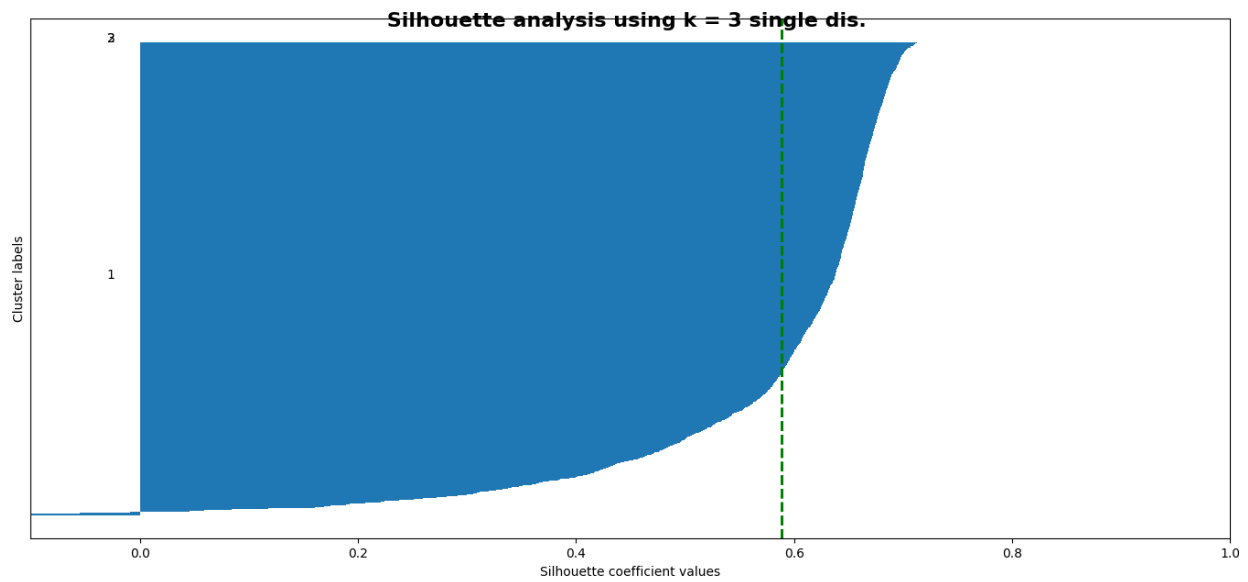


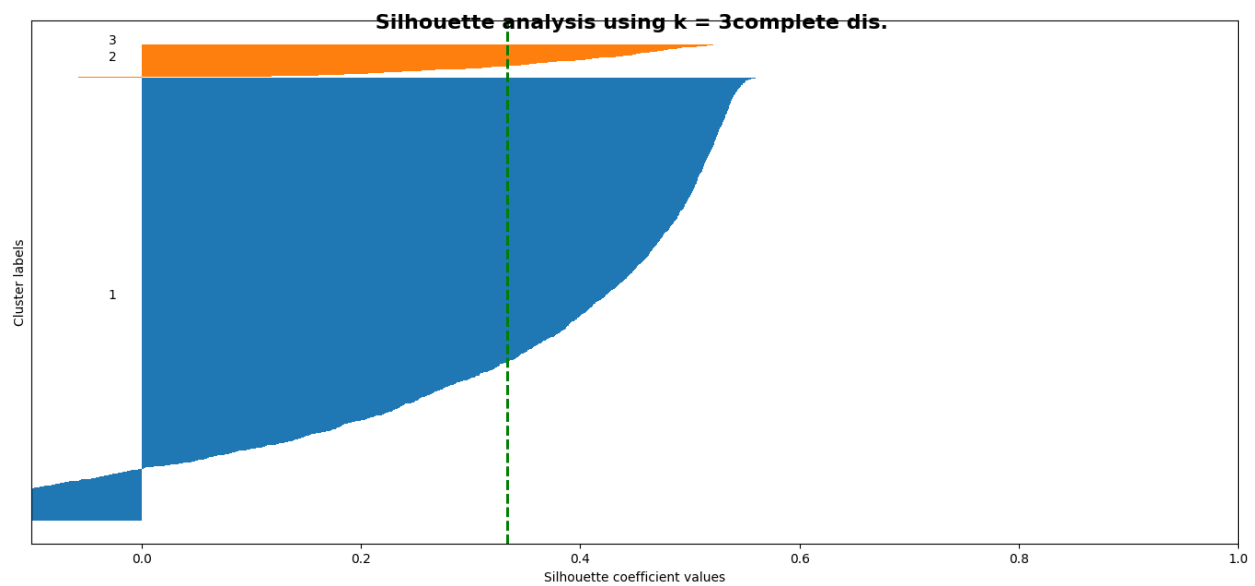
باتوجه به نمودارهای Silhouette بهترین  $k$  برابر با ۲ است؛ زیرا خط نقطه چین به یک نزدیک تر است. اما در این  $k=2$  خلوط خیلی افتضاح است. درنهایت با درنظر گرفتن هردو فاکتور بنظر میرسد  $k=3$  بهتر از بقیه است.

در این قسمت چون تعداد داده‌ها خیلی زیاد بود و عملاً نمیتوانستم از مرحله‌ی یک ادغام شدن داده‌ها را نشان دهم، به صورت سطح بالا نگاه کردم و مراحل ابتدایی را نشان ندادم در نمودارها.

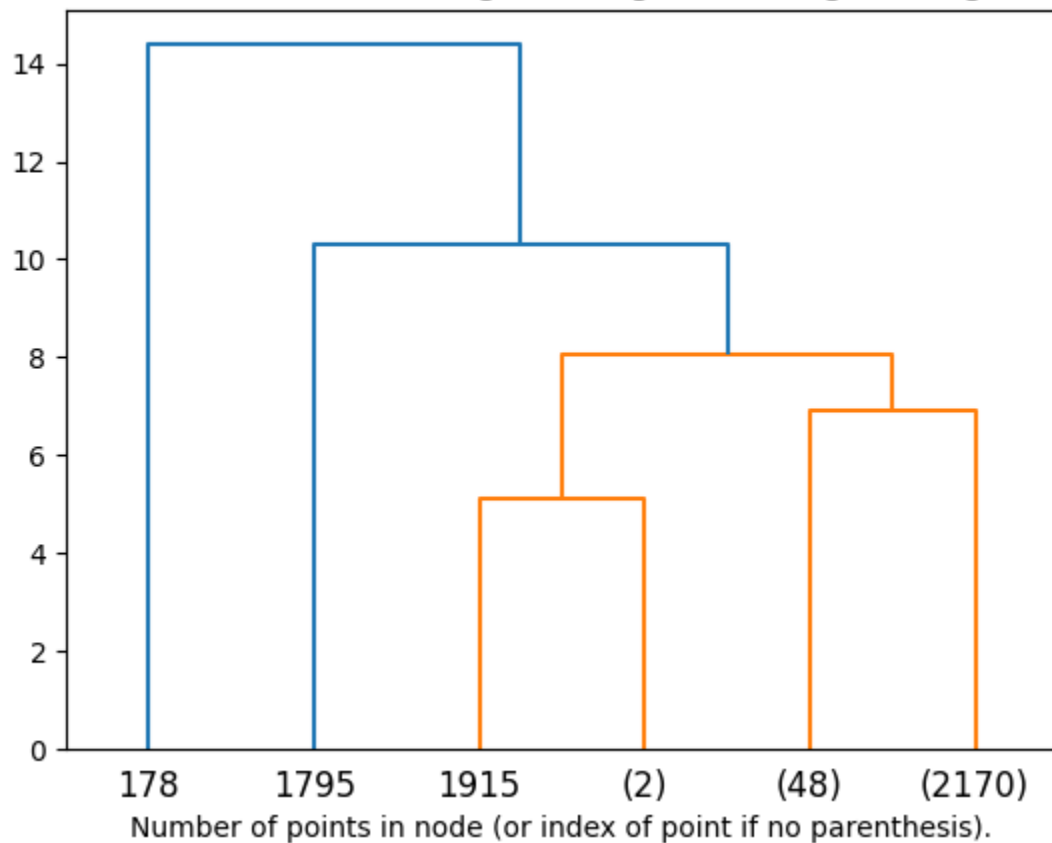
در محور افقی مشخص شده است که در هر شاخه چند داده قرار گرفته اند.

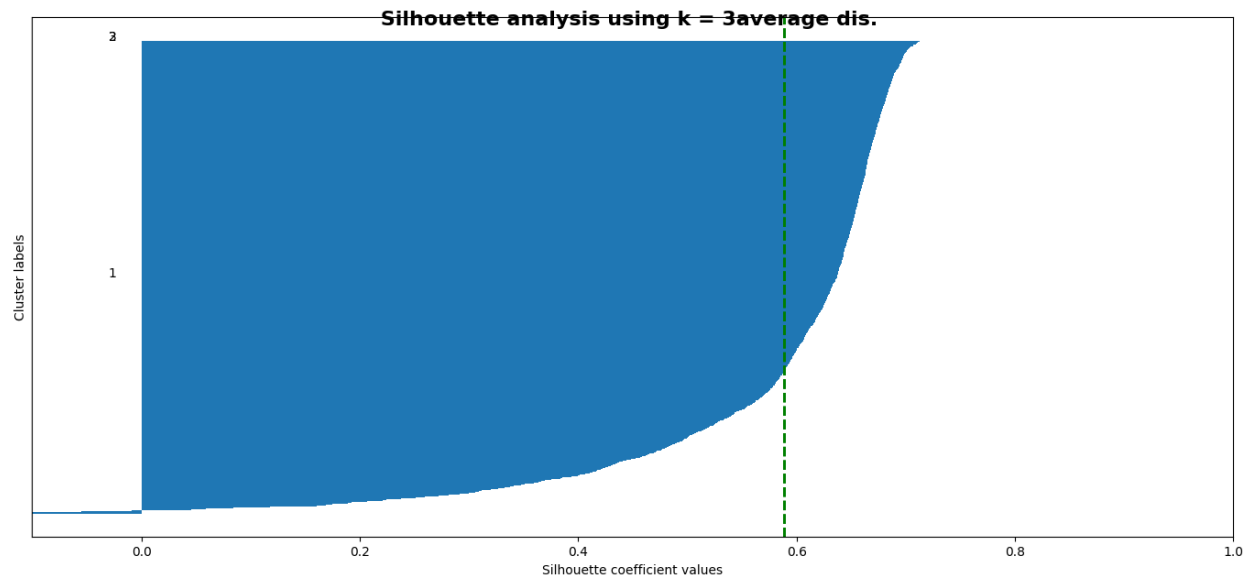






**Hierarchical Clustering Dendrogram Average Linkage**





با دقت به نمودارهای silhouette میبینیم در دو فاصله‌ی average, single خط نقطه چین به یک نزدیک‌تر است. بنابراین این دو فاصله برای این داده‌ها بهتر عمل میکنند تا فاصله‌ی complete.

## سوال هفتم

(a)

در این سوال ابتدا داده‌ها flatten شدند و سپس تمامی داده‌ها نرمال شدند و در بازه‌ی ۱- و ۱ قرار گرفتند. (برای pca ضروری است که میانگین داده‌ها صفر شود ولی در بازه‌ی ۱- و ۱ قرار گرفتن الزامی نیست).

[173.67488588 112.39125963 43.0300822 40.00863069 31.89091725

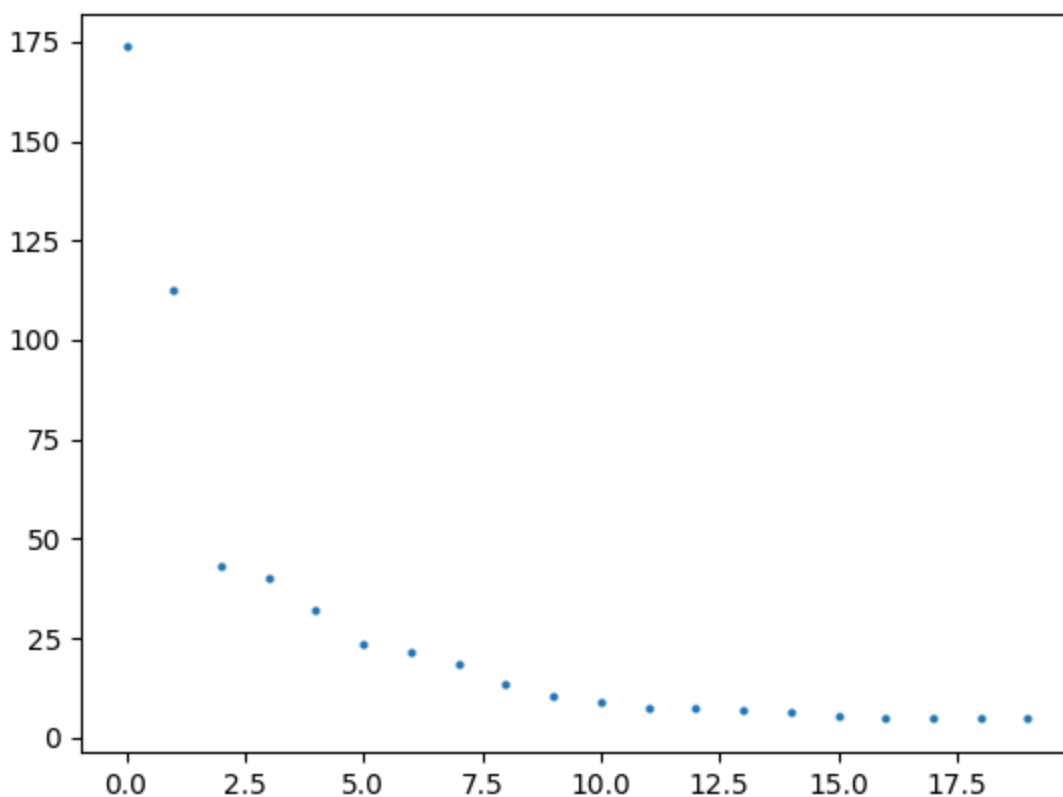
23.47090509 21.79615273 18.36830105 13.68573291 10.74170855

9.12172719 7.70949108 7.32528266 6.77152194 6.53217671

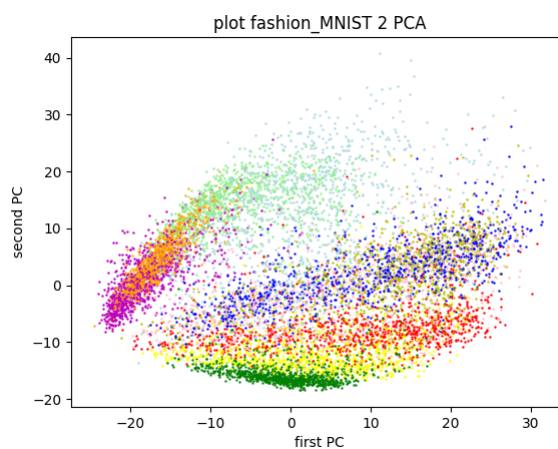
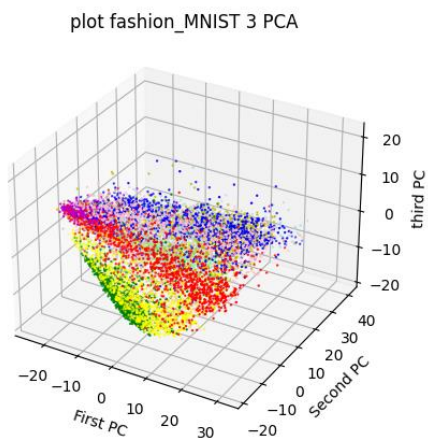
5.66578484 5.14191367 5.01986023 4.90432207 4.80492982]

نمایش مقادیر ویژه به صورت زیر است.





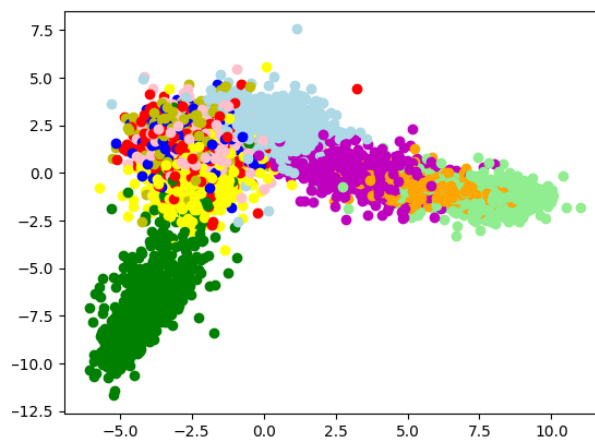
تصویر داده‌ها به دو و سه ویژگی اول به صورت تصاویر زیر است.



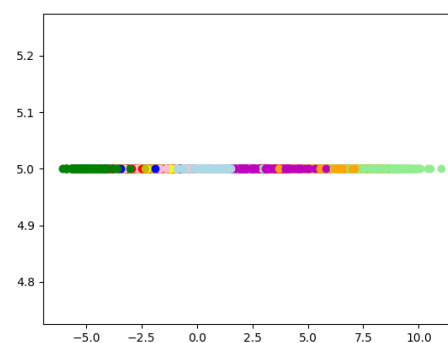
همانطور که مشخص است اگر در تصویر سمت چپ داده‌ها را در بعد سوم بر صفحه‌ی افقی تصویر کنیم تصویر سمت راست حاصل میشود. در هر دو حالت نیز داده‌ها جداپذیر نیستند.

(b)

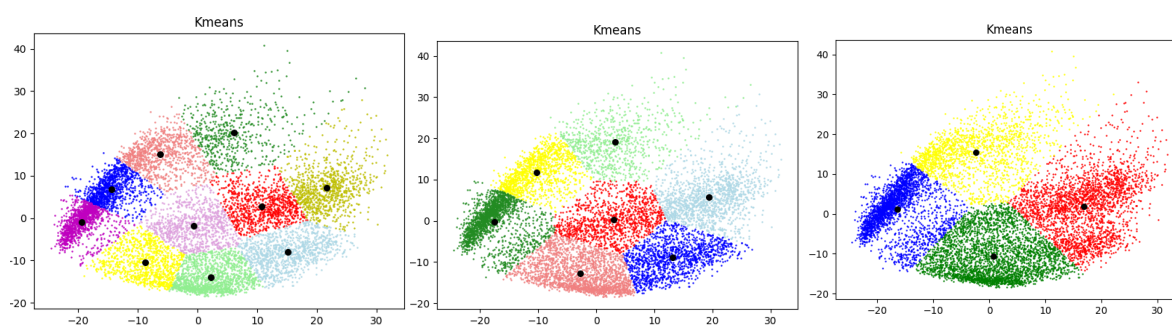
برای دو کامپوننت



برای یک کامپوننت

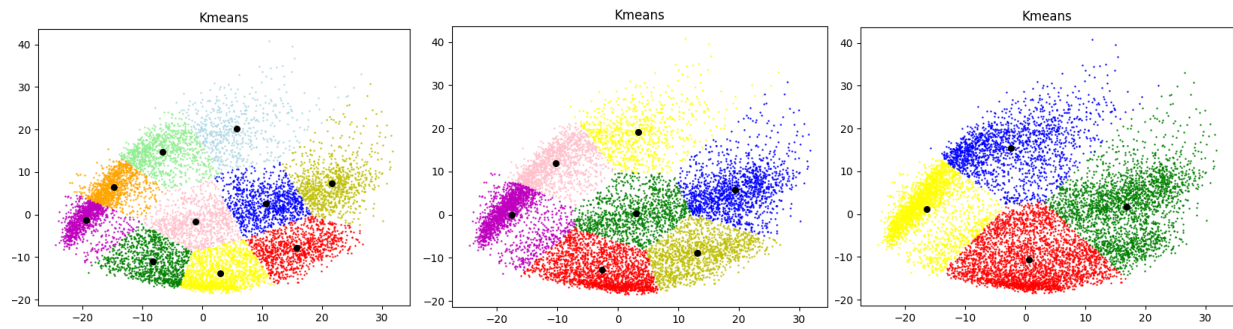


(C)



همانطور که انتظار می‌رفت داده‌ها به شکل صحیحی خوشه‌بندی نشدند. زیرا kmeans داده‌ها را به صورت خطی خوشه‌بندی میکند و سعی میکند خوشه‌ها فشرده و داده‌ها حول مرکز خوشه باشند.

(d)



همانطور که مشاهده میشود در این روش نسبت به روش قبل تفاوت چندانی حاصل نشده است.

یک دلیل این است که k-means خطی جدا میکند و این داده‌ها خطی جداپذیر نیستند با دو بعد.

هر روشی برای انتخاب مراکز در نظر گرفته شود حاصل نهایی همین است.

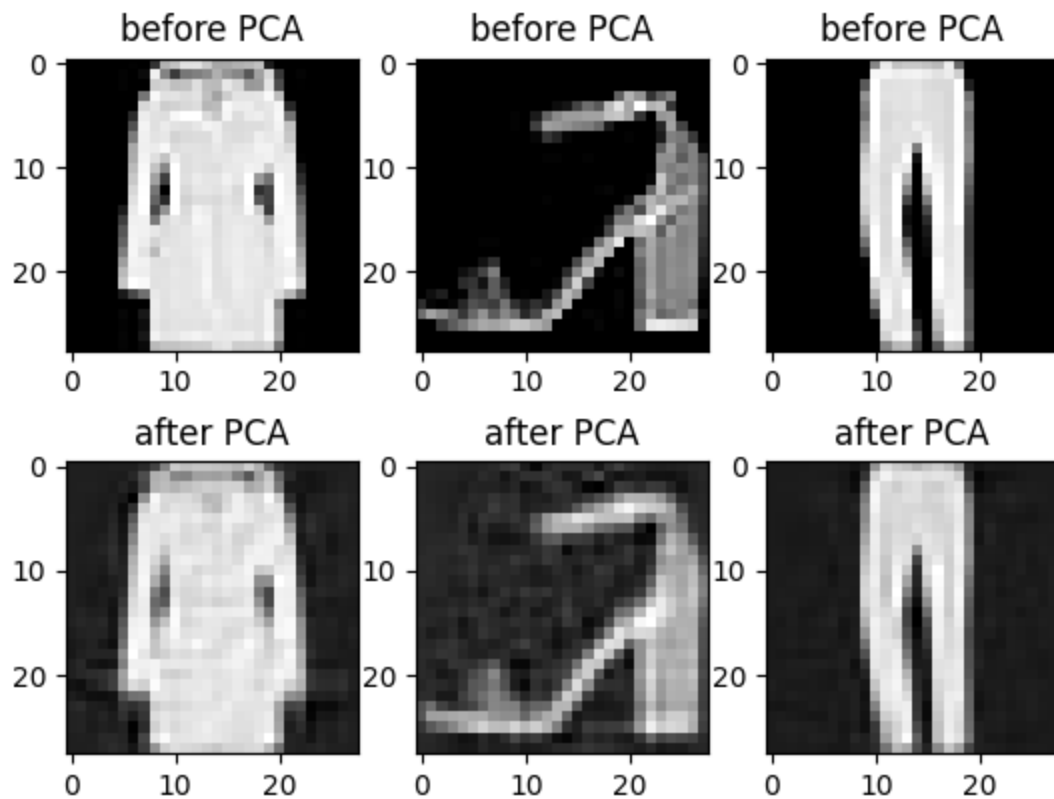
برای رسیدن به خوشه بندی بهتر باید تعداد کامپوننت های بیشتری را انتخاب کرد.

(e)

```
for i in range(2, 748):
    pca = PCA(n_components=i)
    Y = pca.fit(data)
    if sum(pca.explained_variance_ratio_) > 0.95:
        print(i)
        break
```

خروجی: ۱۸۵

۱۸۵ ویژگی اول اگر انتخاب شود دقت مد نظر حاصل می‌شود. (۰.۹۵۰۲)



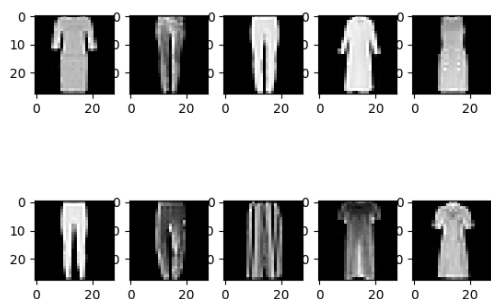
(f) (به اشتباه عنوان تصاویر کلاس نوشته شده، باید مینوشتم خوشه)

خروجی هر خوشه قابل مشاهده است. همانطور که مشهود است در بعضی از خوشه‌ها اشتباه وجود دارد و این به دلیل شبیه بودن ساختار دو جنس متفاوت است. به طور مثال در خوشه‌ی ۱ پیراهن بلند و شلوار ساختاری شبیه به هم داشتند و این اشتباه رخ داده است.

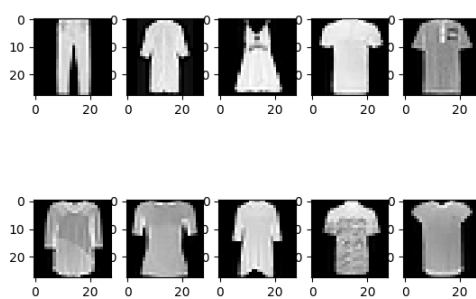
یکی از دلایل رخداد این اتفاق این است که تعداد خوشه‌ها کمتر از تعداد خوشه‌های واقعی است. (خوشه با در نظر گرفتن ساختار؛ یعنی پیراهن بلند با بلوز جدا در نظر گرفته شود، انواع لباس‌های زنانه خوشه‌های مجزا داشته باشند).

اگر تعداد خوشه‌ها بیشتر بود این اتفاق کمتر رخ میداد.

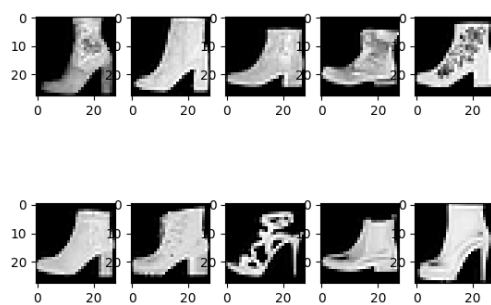
class1



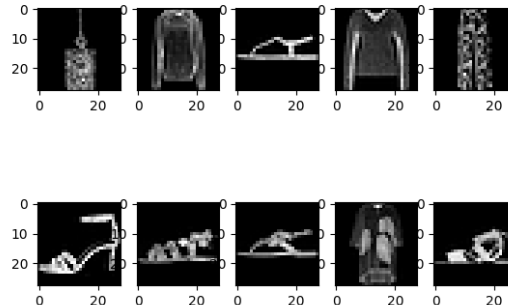
class0



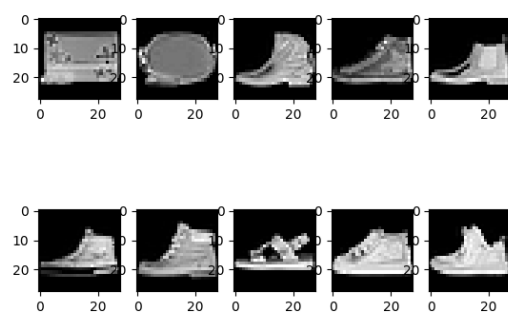
class3



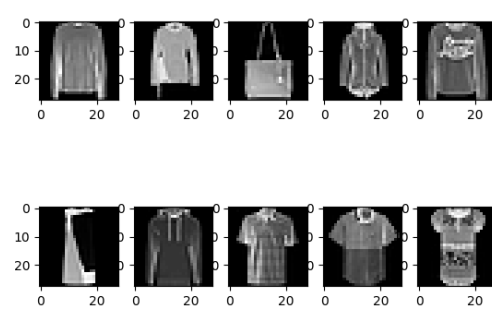
class2



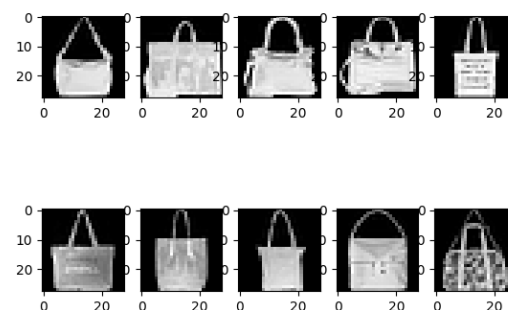
class5



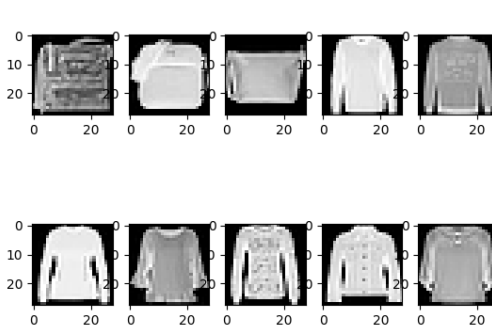
class4

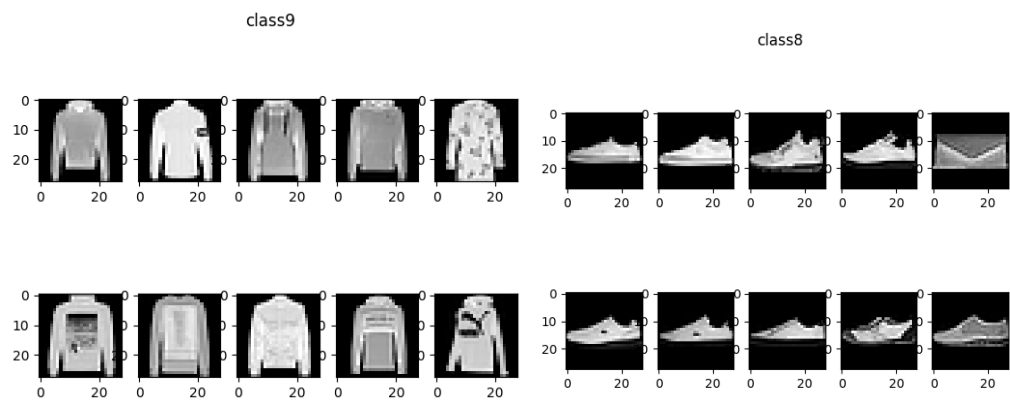


class7

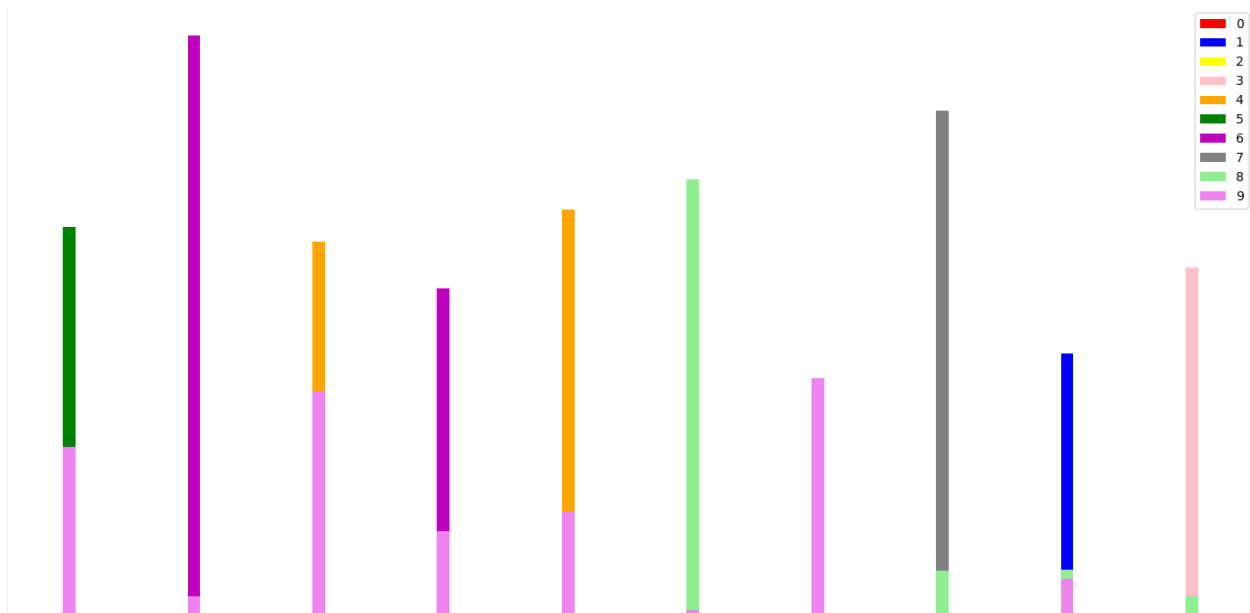


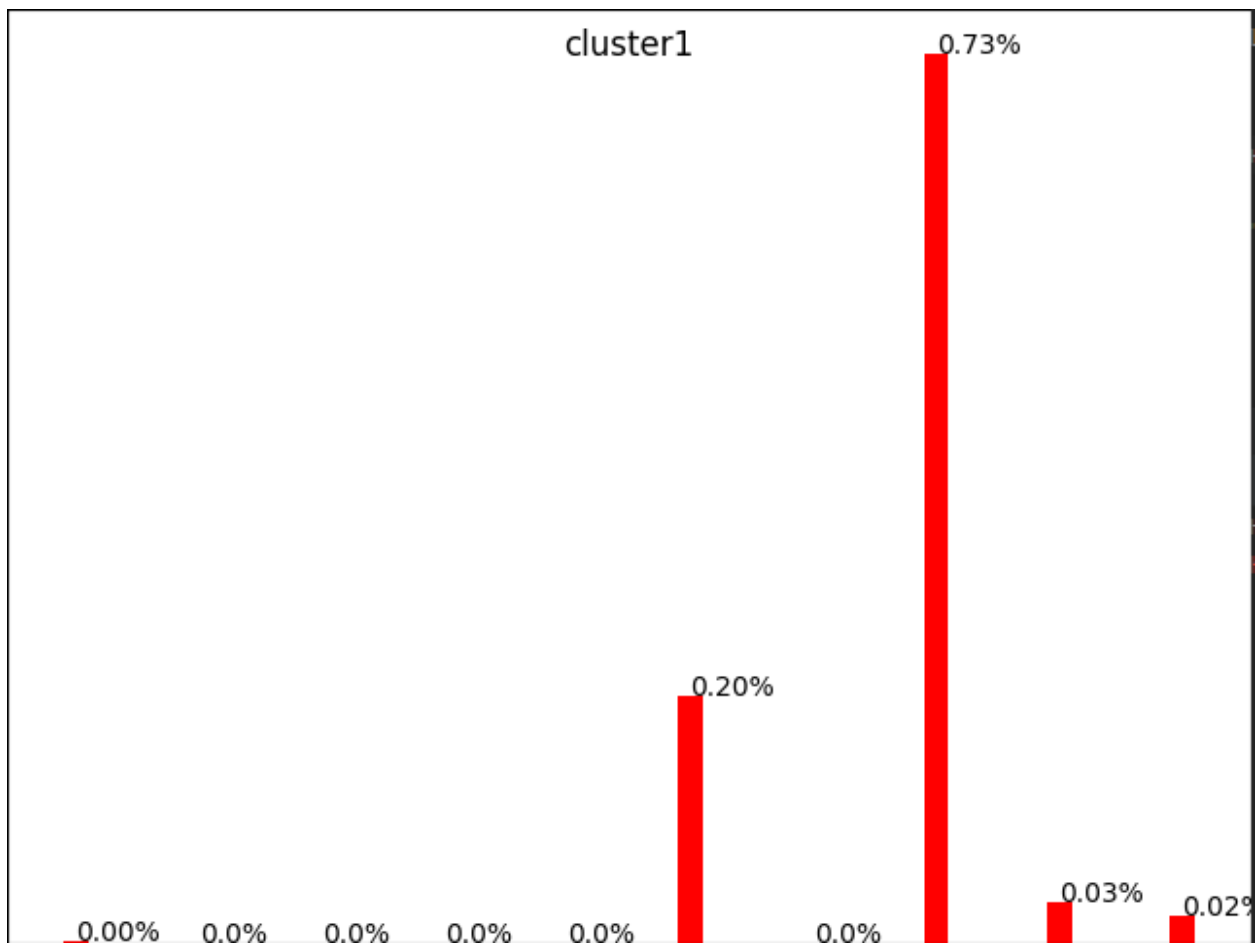
class6

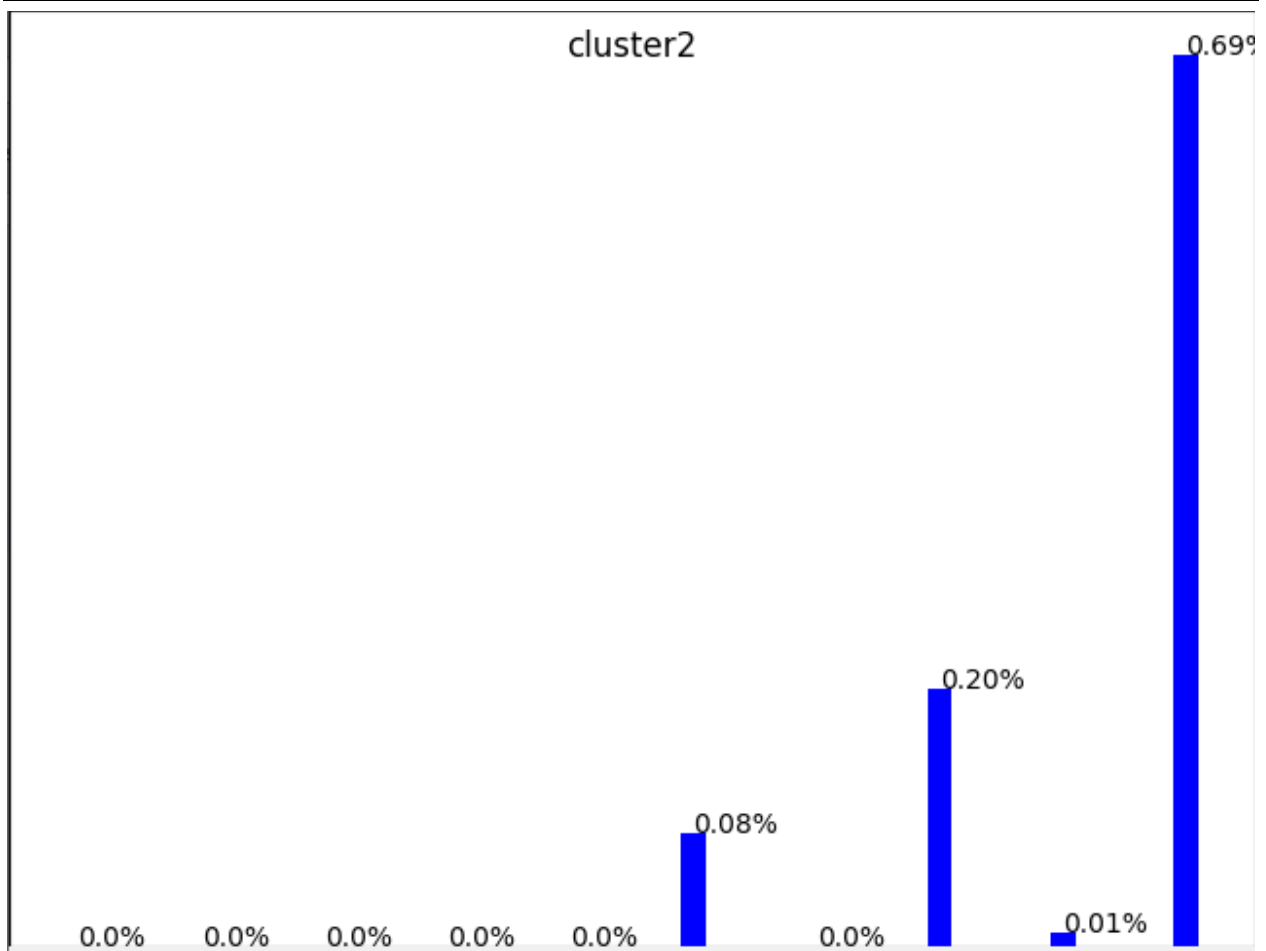




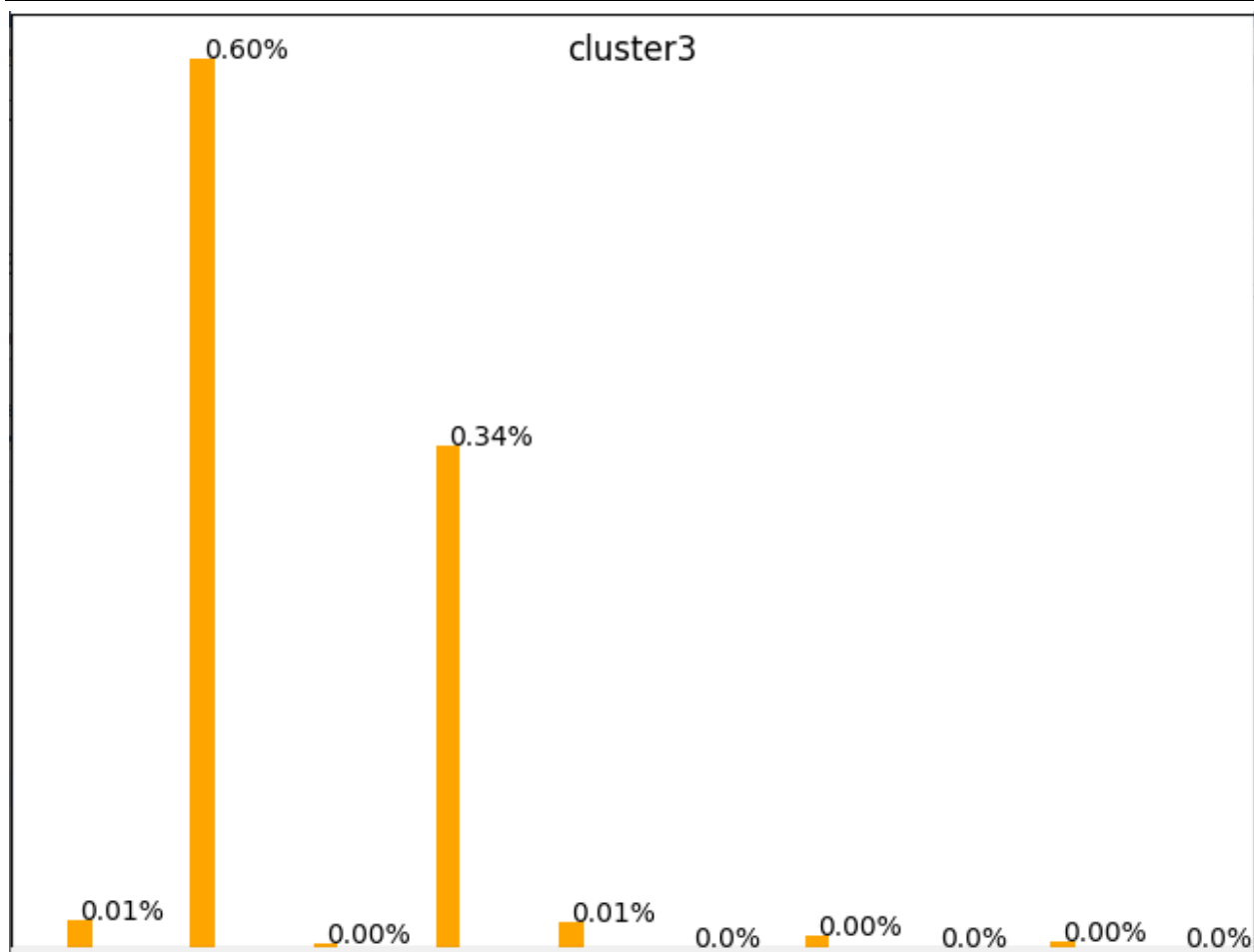
(g)

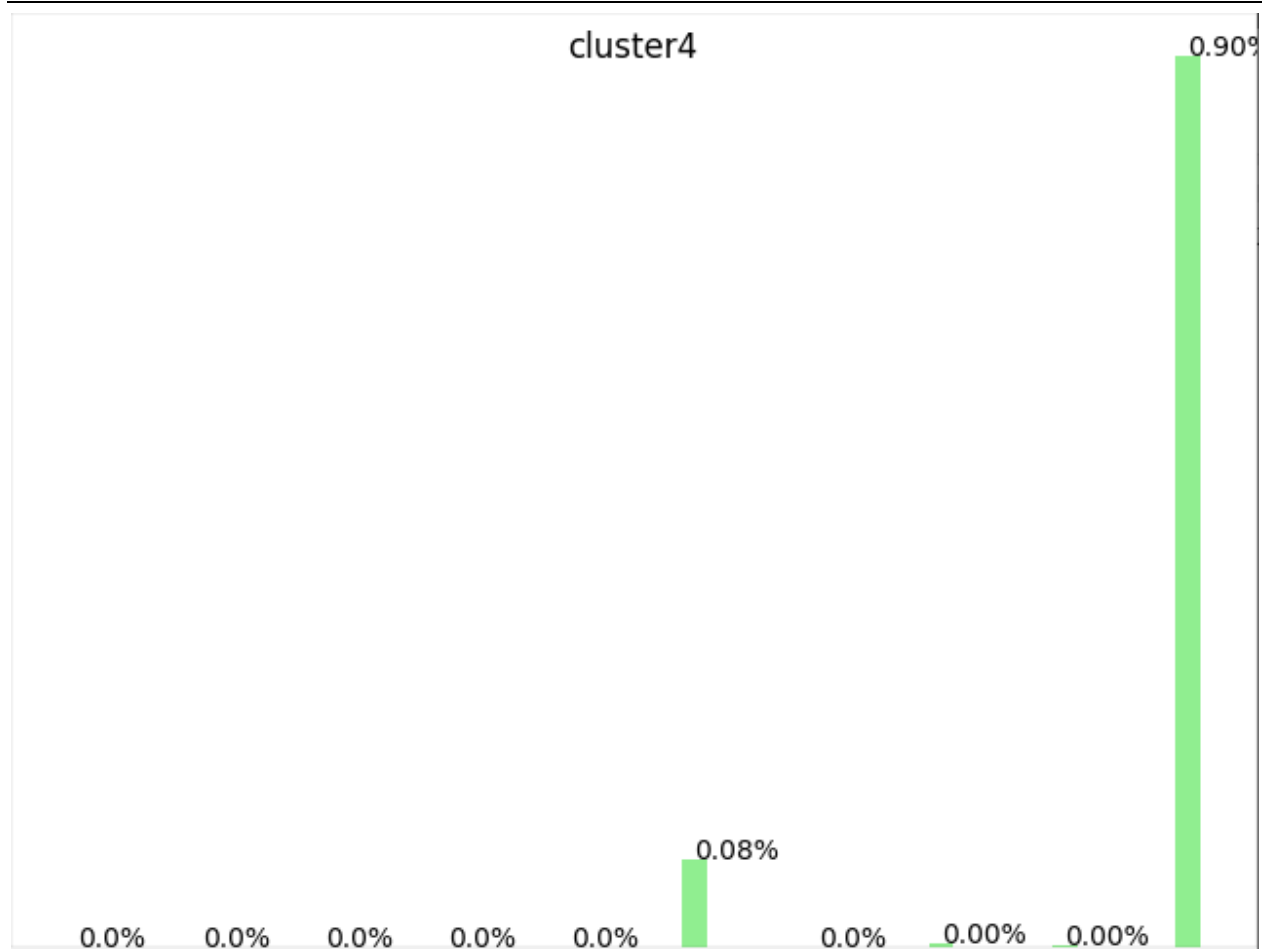


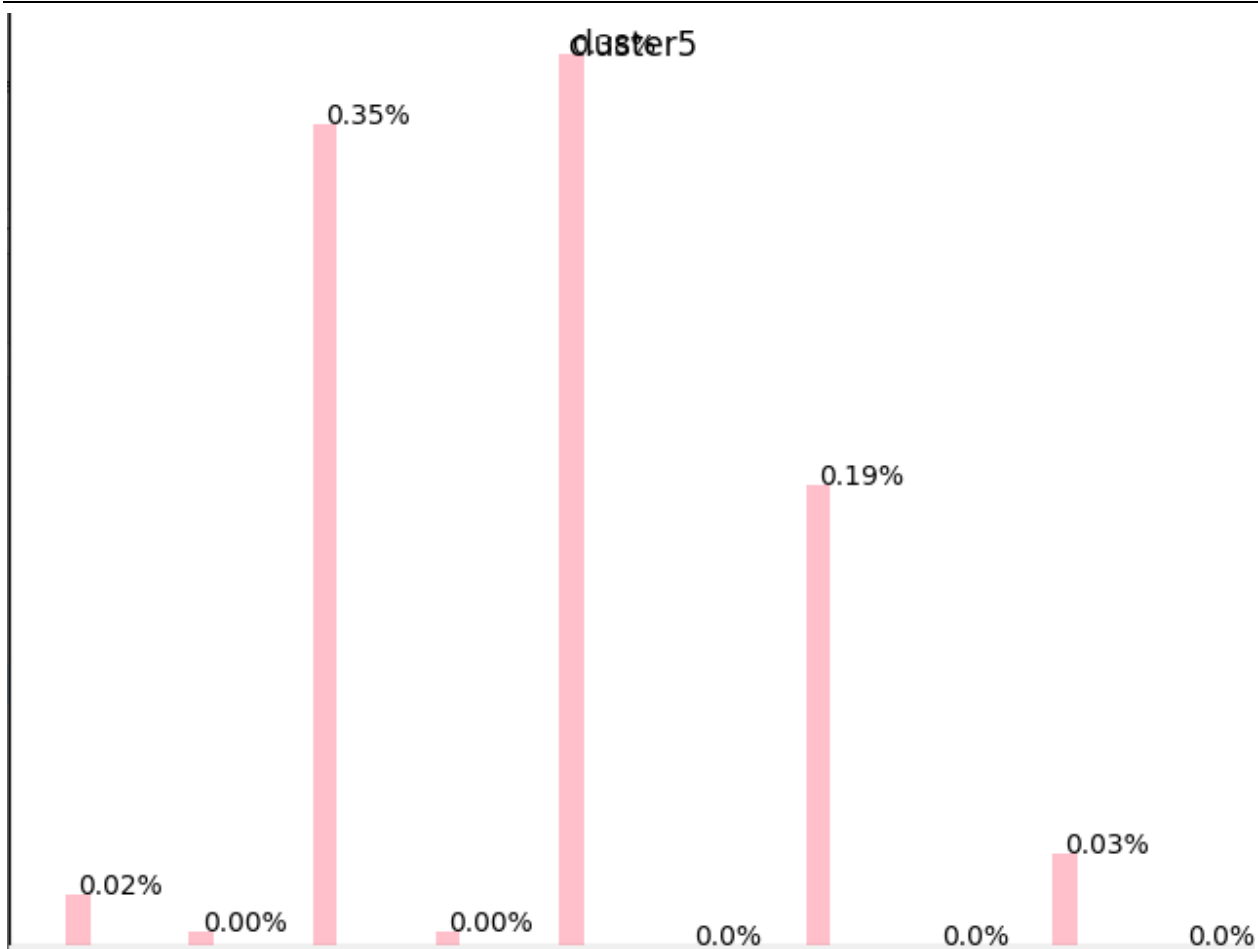


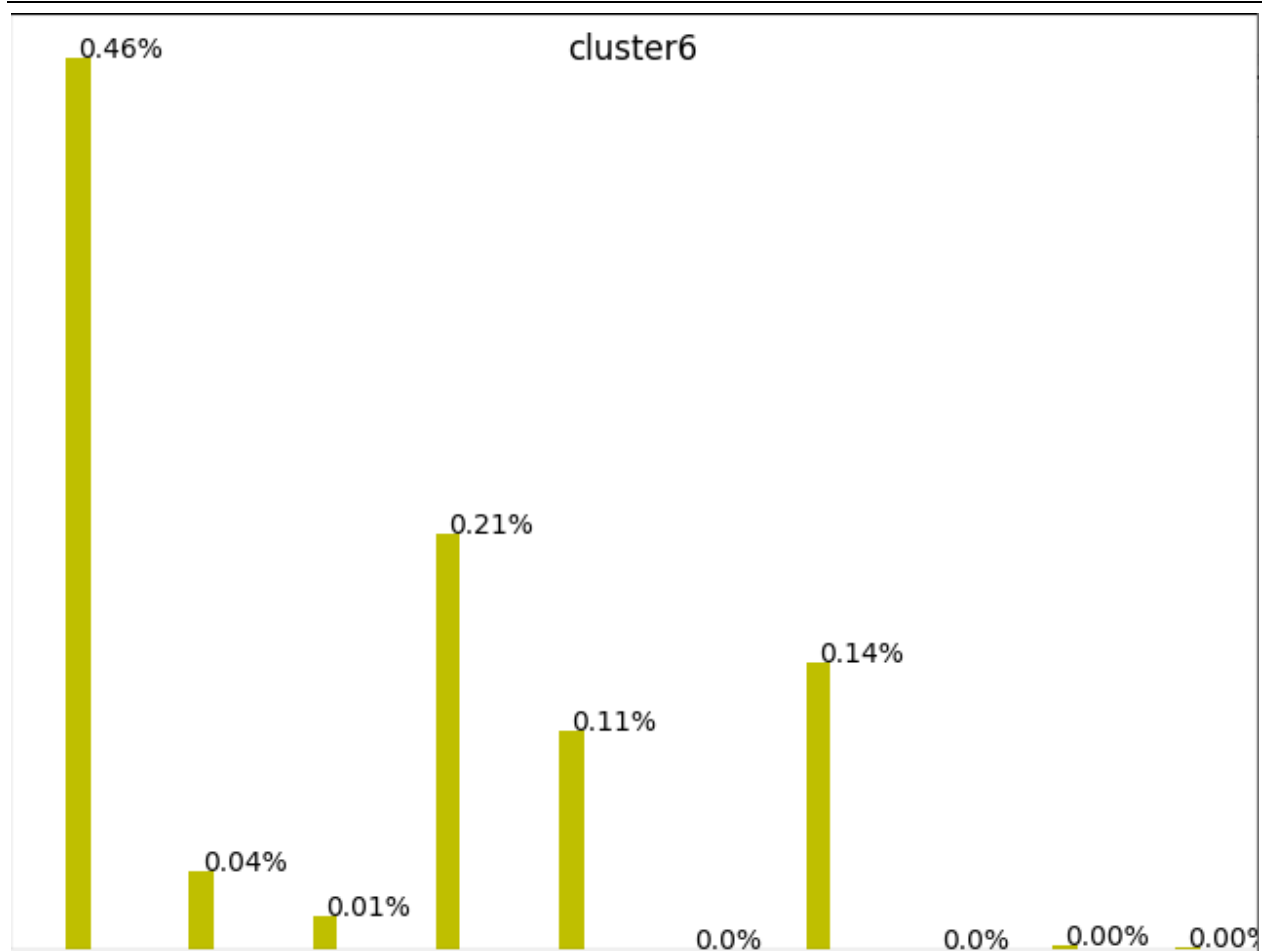


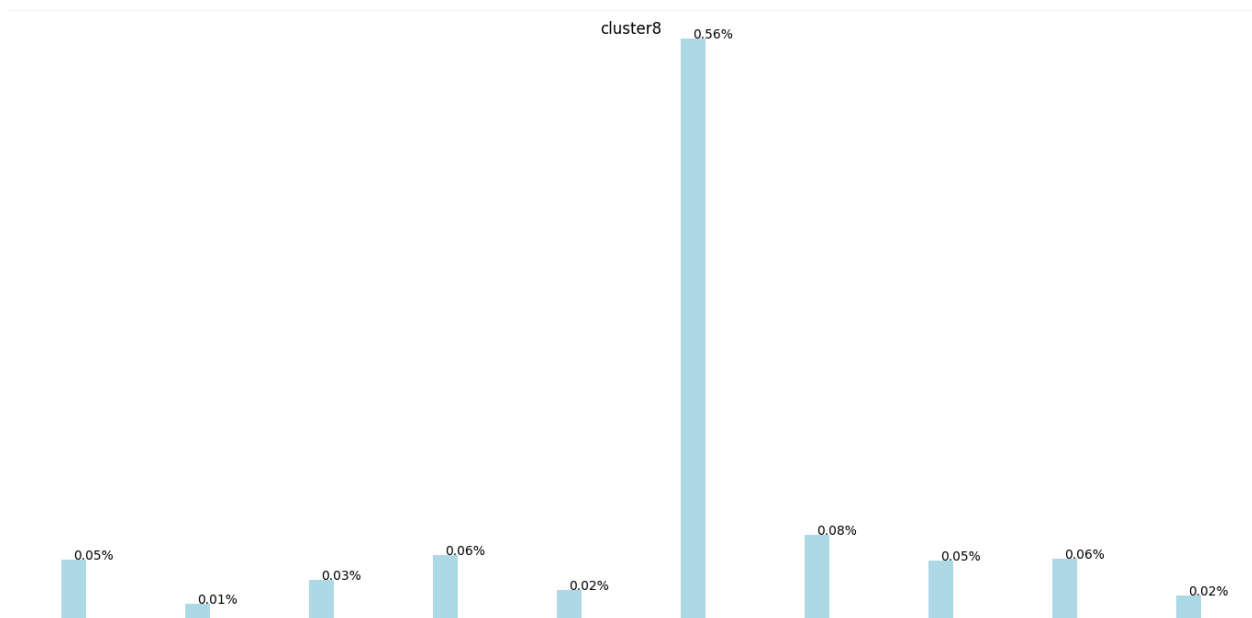
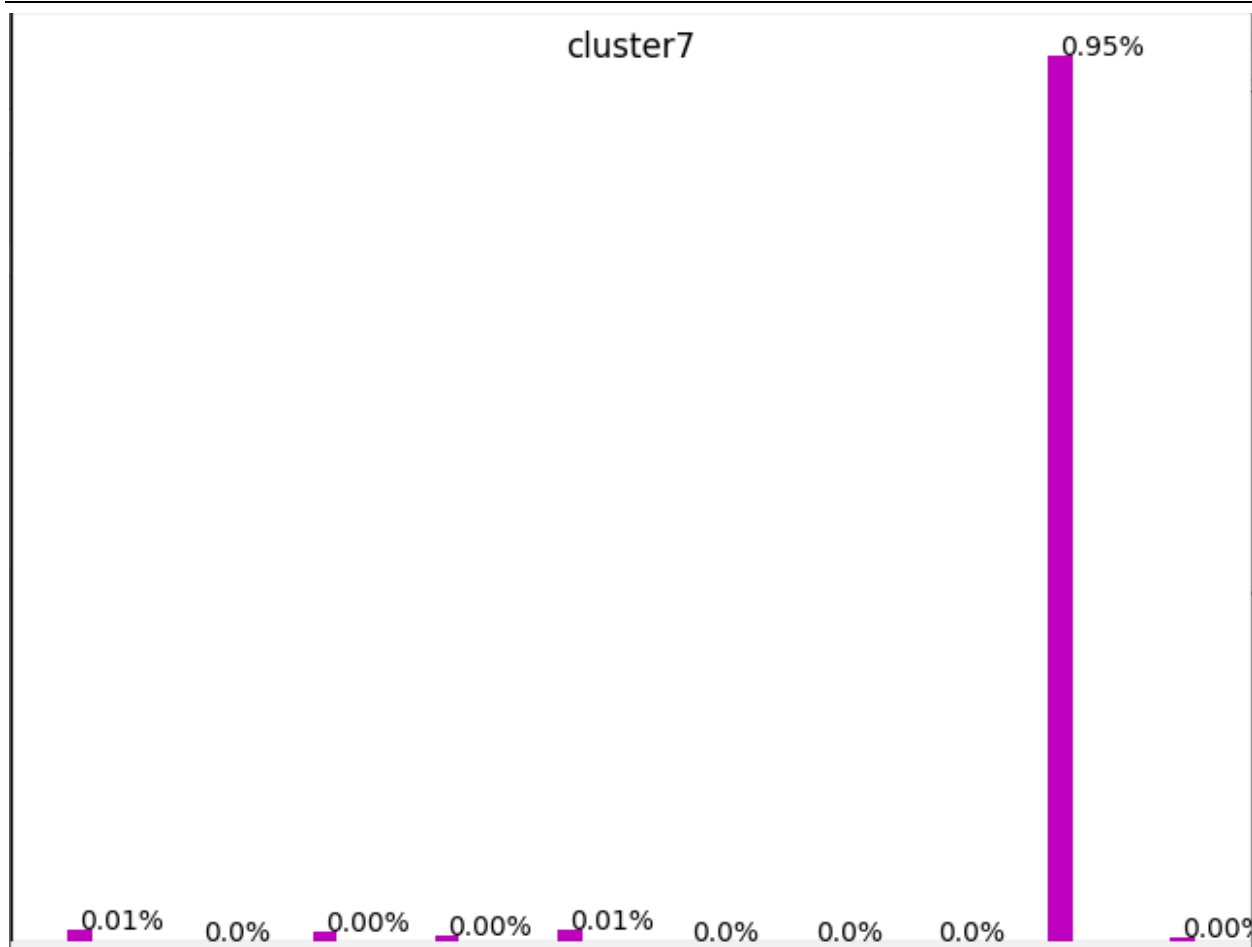


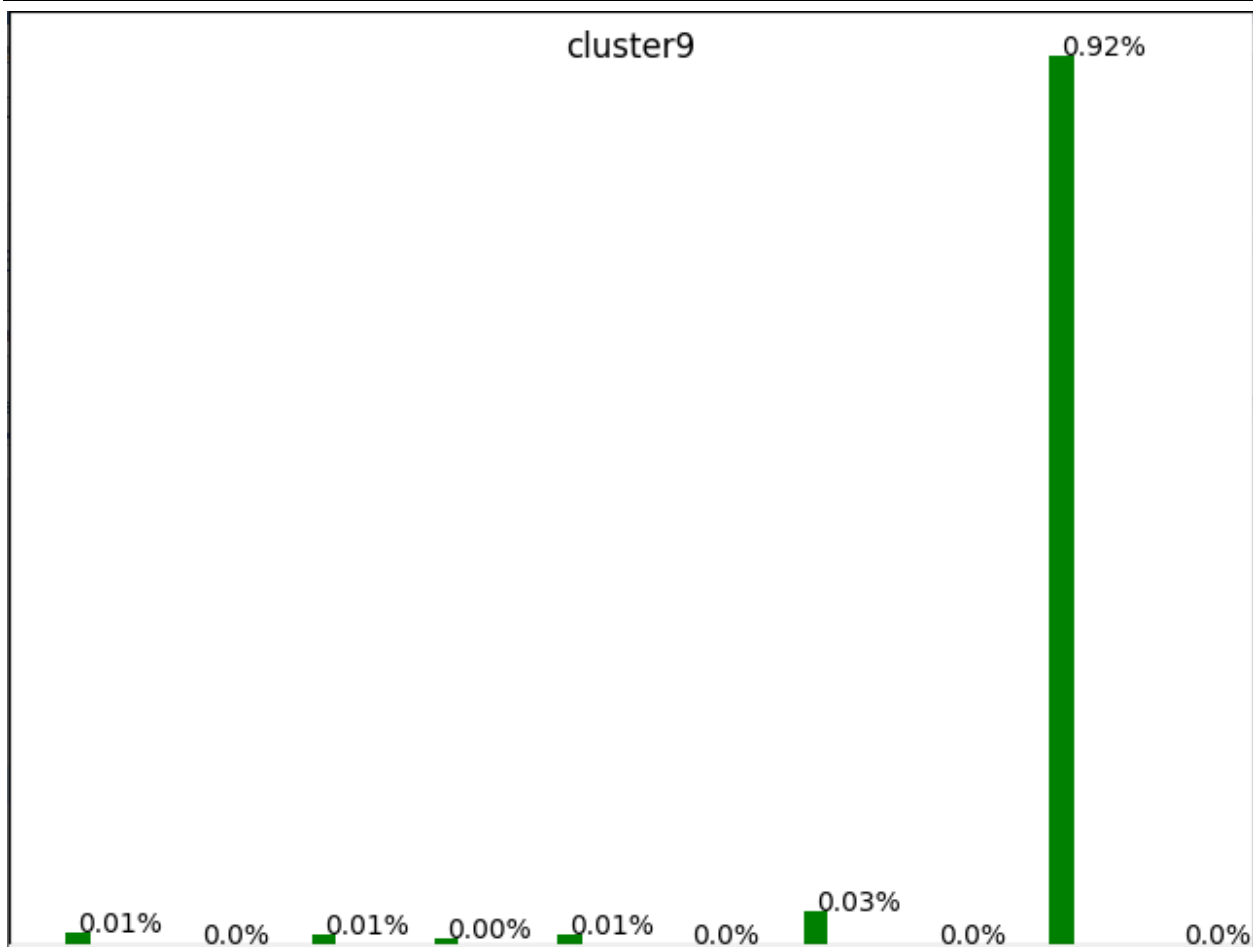


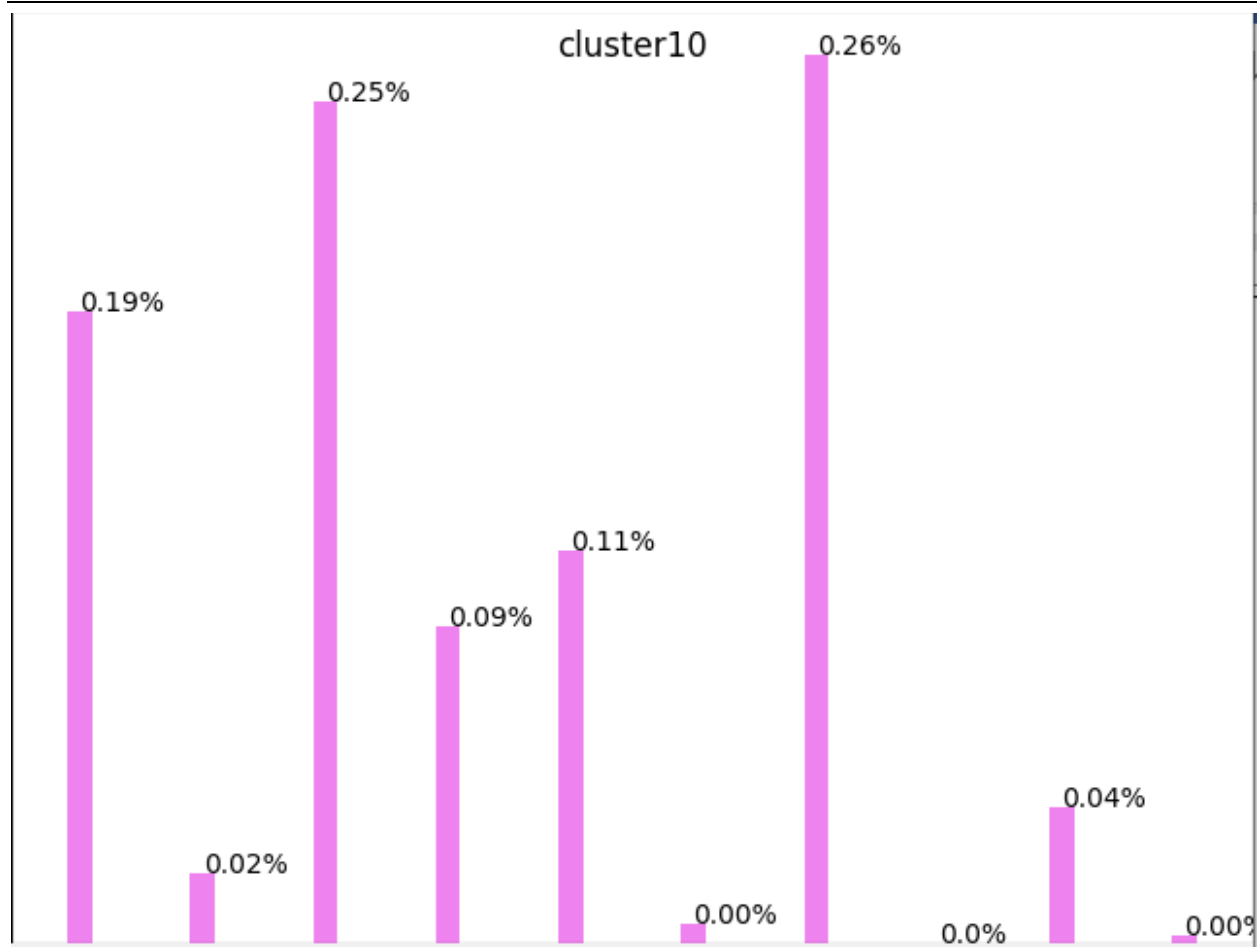






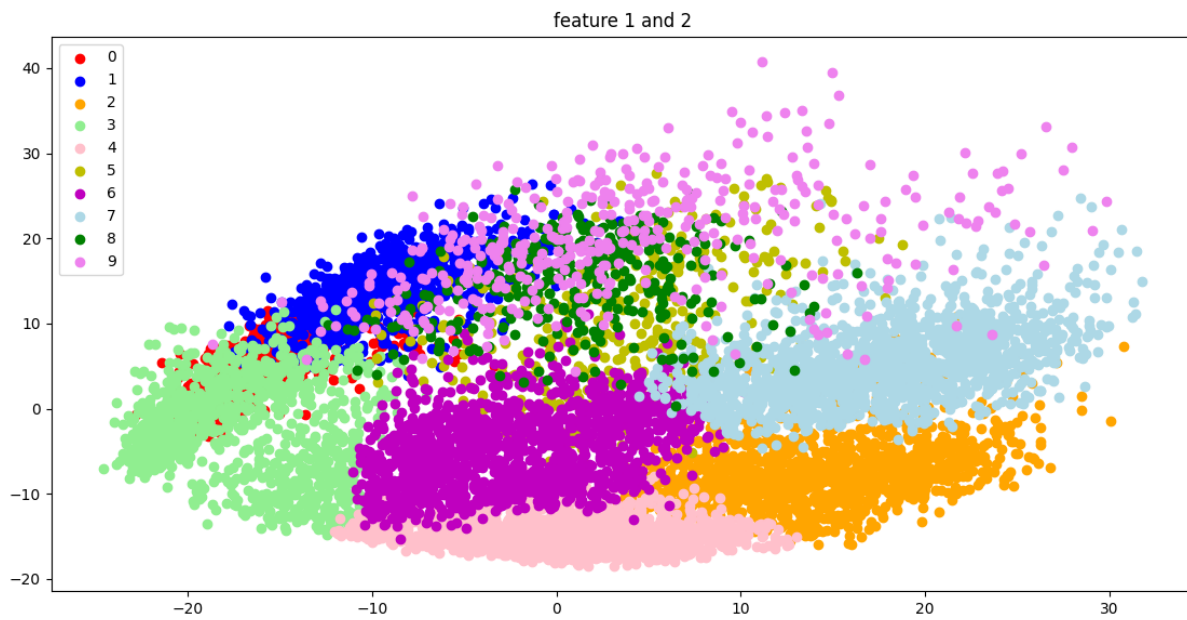




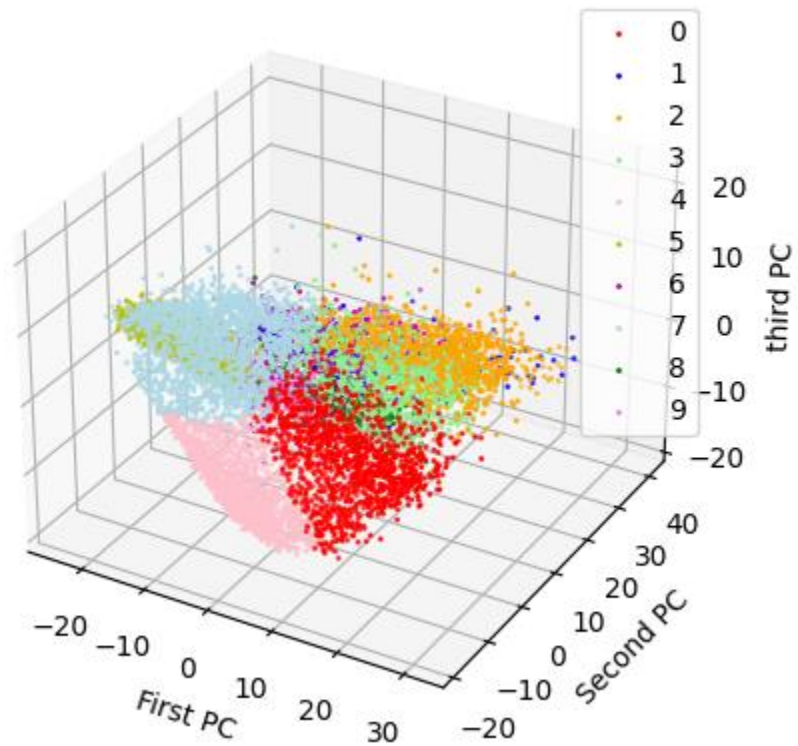


همانطور که از این نمودار ها متوجه شدیم، خلوص خوشه ها یکسان نیست و در بعضی از خوشه ها مانند خوشه ی ده خلوص خیلی کم است. این نتایج مطابق با قسمت قبل است که از هر خوشه ۱۰ داده را نمایش دادیم و در بعضی از خوشه ها داده ها یک جنس نبودند.

(h)



plot fashion\_MNIST 185 PCA and k-means clustering





## سوال هشتم

(a) اگر داده‌های مساله در فضای  $n$  بعدی باشند و تعداد داده ها  $m$  باشد و  $m < n$

انگاه ماتریس داده‌ها حداکثر دارای rank برابر با  $n$  است.

همانطور که میدانیم در  $pca$  از ماتریس کواریانس داده‌ها استفاده می‌شود و این ماتریس  $n \times n$  است.

پس در نهایت ما میتوانیم  $n$  تا مقدار ویژه با استفاده از این ماتریس بدست آوریم چون rank ماتریس کواریانس داده‌ها  $n$  است.

با توجه به این توضیحات  $pc$  ها که متناظر با مقادیر ویژه هستند نمیتوانند بیشتر از  $n$  باشند.

(b)  $pca$  با کاهش ابعاد داده‌ها به فشرده‌سازی کمک می‌کند. اما برای محاسبه‌ی  $pca$  در تصاویر به روش سابق عمل نمیکنیم. زیرا در فضای تصاویر ابعاد خیلی بیشتر از تعداد داده‌ها هستن (مثلا حدود ۲۰۰۰۰۰۰ بعد). طبیعتا یافتن ماتریس  $scatter$  برای چنین بعدی خیلی زمان‌بر است. از یک راه میانبر استفاده میشود و ماتریس  $scatter$  با بعد تعداد داده‌ها ساخته می‌شود. به صورت زیر:

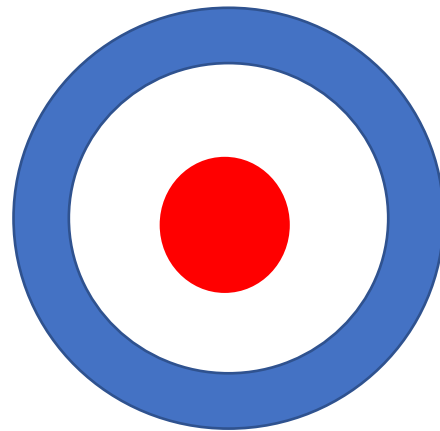
$$A = [ (x_1 - \mu) \dots (x_n - \mu) ]$$

$$S = (n-1) AA^T$$

از این جا به بعد دیگر محاسبات مانند روش معمول است.

(c) بله ممکن است. بعد داده‌ها را به نحوی افزایش می‌دهیم که اجرای الگوریتم  $k$ -means امکان پذیر باشد. به طور مثال اگر توزیع داده‌ها به صورت زیر باشد در فضای دو بعدی، خوشه قرمز و خوشه‌ی آبی، با الگوریتم  $k$ -means نمیتوان در فضای دوبعدی این داده‌ها رو جدا کرد. اما اگر داده‌ها را به فضای سه بعدی ببریم؛ یعنی داده‌های قرمز در ارتفاع متفاوتی با داده‌های خوشه‌ی آبی قرار بگیرند، این الگوریتم کار می‌کند و بردن داده‌ها به فضای مناسب سه بعدی با کرنل انجام می‌شود.

مزیت کرنل این است در مواقعی که داده‌ها خطی جداپذیر نیستند، داده‌ها را به فضایی میبرد که خطی جداپذیر باشند.



(d) با افزایش  $k$  میزان واریانس نیز افزایش می‌یابد. زیرا هرچه  $k$  بیشتر می‌شود تعداد خوشه‌ها بیشتر و همینطور تعداد مراکز بیشتر.

همانطور که میدانیم مراکز نهایی خوشه بندی  $k$ -means وابستگی زیادی به انتخاب اولیه دارد و اگر تعداد  $k$  بیشتر باشد تعداد حالت های ممکن برای انتخاب مراکز بیشتر میشود و در نهایت نیز خوشه بندی های متفاوتی حاصل میشود که به میزان بیشتری به داده ها وابسته هستند؛ یعنی واریانس خوشه بندی بیشتر می شود.

تعریف من از واریانس این است که به تعداد  $m$  بار داده ها را خوشه بندی کنیم و در نهایت واریانس جواب ها را بگیریم.

(e) با توجه به تعریف ذکر شده در قسمت قبل، واریانس صفر زمانی حاصل می شود که  $k=1$  باشد. در این حالت هرچند بار که از الگوریتم اجرا بگیریم خوشه بندی نهایی به یک شکل است و واریانس صفر میشود.