**به نام خدا**

**مرضیه علیدادی – ۸۱۰۱۰۱۲۳۶ تکلیف عملی ۳ داده کاوی**

---

## ❖ سوالات عملی:

### سوال ۱:

- **پیش پردازش:**

ابتدا وجود مقادیر گمشده در مجموعه داده‌ی موردنظر، بررسی شد:

| | Total | Percent |
|---|---|---|
| OCCUPATION | 665 | 0.064551 |
| CAR_AGE | 639 | 0.062027 |
| HOME_VAL | 575 | 0.055814 |
| INCOME | 570 | 0.055329 |
| YOJ | 548 | 0.053194 |
| AGE | 7 | 0.000679 |
| ID | 0 | 0.000000 |
| TIF | 0 | 0.000000 |
| CLAIM_FLAG | 0 | 0.000000 |
| CLM_AMT | 0 | 0.000000 |
| MVR_PTS | 0 | 0.000000 |
| REVOKED | 0 | 0.000000 |
| CLM_FREQ | 0 | 0.000000 |
| OLDCLAIM | 0 | 0.000000 |
| RED_CAR | 0 | 0.000000 |
| CAR_TYPE | 0 | 0.000000 |
| TRAVTIME | 0 | 0.000000 |
| BLUEBOOK | 0 | 0.000000 |
| CAR_USE | 0 | 0.000000 |
| KIDSDRIV | 0 | 0.000000 |
| EDUCATION | 0 | 0.000000 |
| GENDER | 0 | 0.000000 |
| MSTATUS | 0 | 0.000000 |
| PARENT1 | 0 | 0.000000 |
| HOMEKIDS | 0 | 0.000000 |
| BIRTH | 0 | 0.000000 |
| URBANICITY | 0 | 0.000000 |

۶ مورد از ستون‌ها، دارای مقادیر گمشده هستند.

ستون‌ها از جهت عدم وجود مقادیر غیرمرتبط، بررسی شدند. برای مثال، ستون‌های categorical، از نظر عدم وجود مقادیری خارج از دسته‌های موجود، بررسی شدند. ستونی از مجموعه داده‌ها دارای مقادیر غیرمرتبط نبود.

مجموعه داده، از نظر outlierها بررسی شد. برای این کار ابتدا ستون‌هایی که قابل تبدیل به نوع عددی بودند (مثلا ستون‌هایی که از جنس رشته بودند و نشان‌گر قیمت دلار بودند)، به داده‌های عددی تبدیل شدند. برای این کار، یک بار با استفاده از روش inter quantile range، outlierهای احتمالی یافت شدند. و بار دیگر

با استفاده از روش z-score، این کار تکرار شد. با بررسی مقادیر یافت شده به عنوان outlierهای احتمالی، این نتیجه حاصل شد که این مقادیر، outlier نیستند و نیازی به حذف آن‌ها نیست.

به داده‌های گمشده، بدین شکل رسیدگی شد:

ستون age که قابل محاسبه از طریق ستون birth بود، مقادیر گمشده‌اش از طریق این محاسبه، بدست آمد.

برای دو ستون occupation و education، مقدار chi2 محاسبه شد و مقدار p-value برابر ۰ بدست آمد، که نشان‌گر وجود رابطه بین مقادیر این دو ستون است. در نتیجه، از ستون education که مقدار گمشده‌ای نداشت، برای پیشبینی مقادیر گمشده‌ی occupation با استفاده از آموزش یک مدل logistic regression استفاده شد.

برای بقیه‌ی ستون‌های دارای مقادیر گمشده، رابطه‌ای برای پیشبینی یا محاسبه‌ی آن‌ها از طریق باقی ستون‌ها یافت نشد، و با توجه به این‌که تعداد گمشده‌های آن‌ها زیاد بود و قابل حذف نبودند، از میانگین هر یک، به عنوان جایگزین مقادیر گمشده‌ی آن‌ها، استفاده شد.

- **استفاده از معیار سیلوئت برای تعیین تغییرات دقت خوشه‌بندی نسبت به تعداد خوشه‌ها در k-means:**

با توجه به این‌که ستون id برای خوشه بندی، اثر مثبتی ندارد، از مجموعه داده‌ی موردنظر حذف شد.

با توجه به این‌که دو ستون birth و age دارای یه معنی هستند و استفاده از ستون age به دلیل این‌که داده‌ی عددی است، ساده‌تر است، ستون birth حذف شد.

داده‌های categorical به داده‌های عددی تبدیل شدند، تا قابل استفاده در الگوریتم k-means شوند.

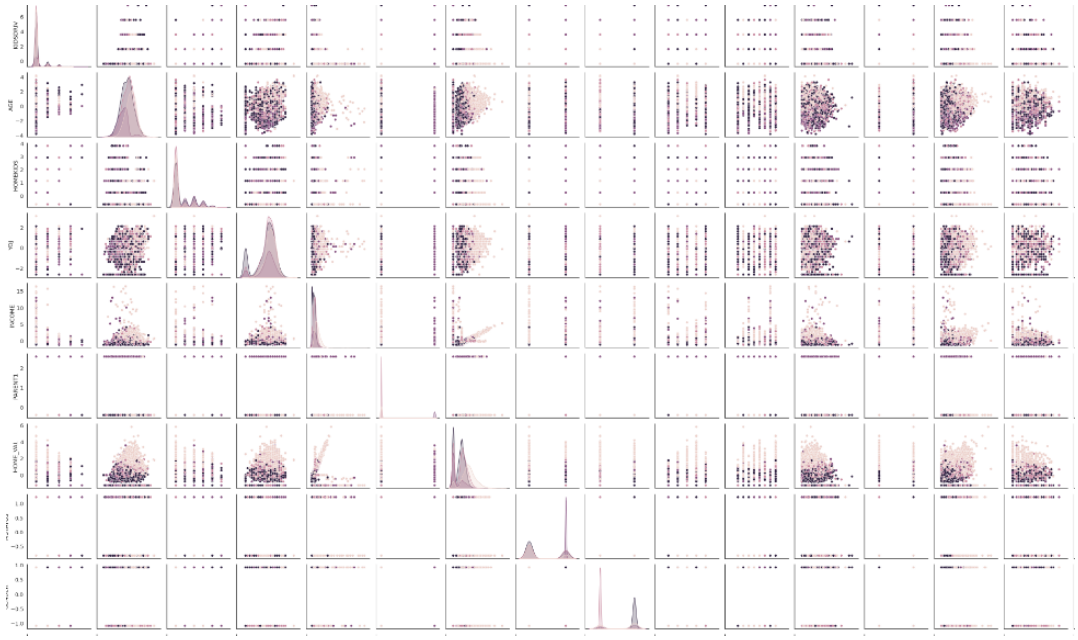
سپس مجموعه داده‌ی موردنظر، نرمال شد. و الگوریتم k-means با قرارگیری مقدار k (که بیان‌گر تعداد خوشه‌ها در خوشه بندی است)، در رنج مقادیر ۲ تا ۵۰، روی این مجموعه داده‌ی حاصل، اعمال شد.

با استفاده از معیار silhouette، دقت هر یک از این خوشه بندی‌ها محاسبه شد. هر چه این مقدار برای خوشه بندی‌ای بیشتر باشد، آن خوشه بندی، دقت بیشتری دارد. در حالتی که k=4 قرار داده شد، بهترین مقدار silhouette، برابر با ۰,۰۹ حاصل شد.

- **اعمال الگوریتم k-means با k=4 بر روی مجموعه داده‌ی موردنظر:**

نمودار pair plot برای هر جفت ستون‌های مجموعه داده‌ها، با نشان دادن هر یک از خوشه‌ها در رنگ‌های متفاوت، رسم شد.

بخشی از این نمودار:



برای بررسی بهتر، correlation هر یک از ستون‌ها با نتیجه‌ی خوشه بندی، با استفاده از دو روش correlation coefficient و chi2 محاسبه شد:

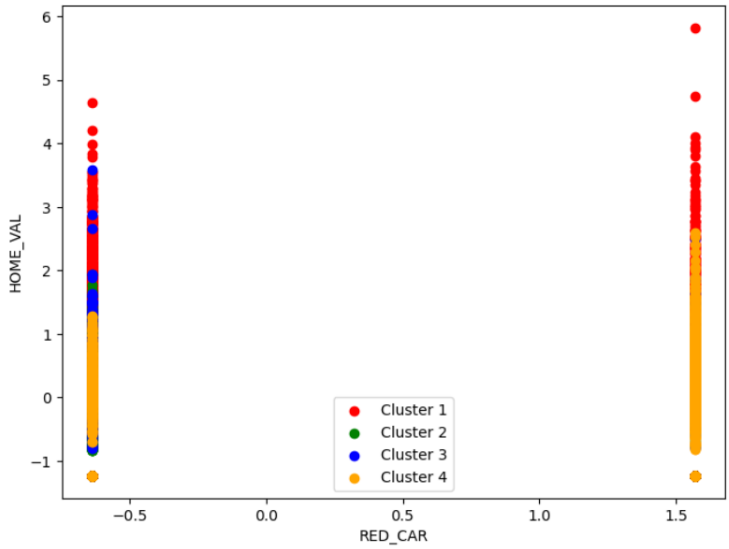
قدر مطلق Correlation coefficientها:

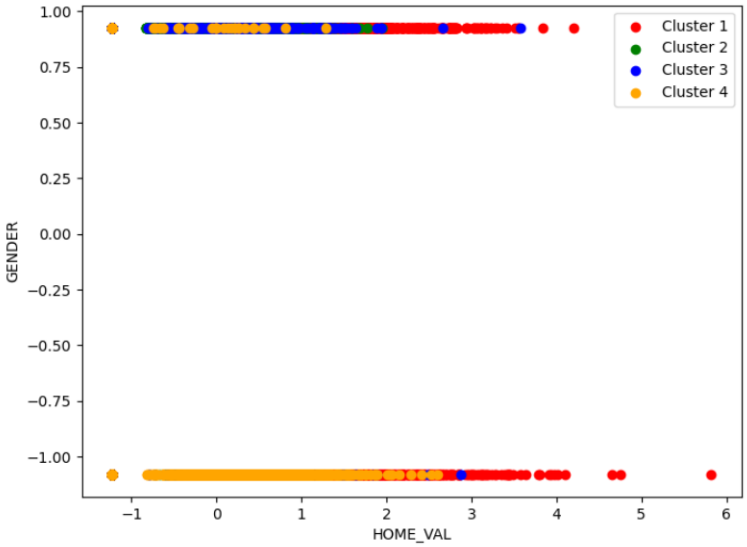
```
cluster_labels    1.000000
RED_CAR           0.581464
GENDER            0.556823
HOME_VAL          0.334302
INCOME            0.251057
CAR_TYPE          0.247825
CAR_AGE           0.236754
BLUEBOOK          0.213055
CAR_USE           0.197356
MSTATUS           0.187418
PARENT1           0.161052
AGE               0.149724
CLAIM_FLAG        0.149241
CLM_FREQ          0.122861
OCCUPATION        0.098648
MVR_PTS           0.097914
HOMEKIDS          0.072343
YOJ               0.051527
URBANICITY        0.048465
KIDSDRIV          0.048286
TRAVTIME          0.042200
REVOKED           0.041689
OLDCLAIM          0.037386
TIF               0.034223
EDUCATION         0.023583
CLM_AMT           0.017123
Name: cluster_labels, dtype: float64
```

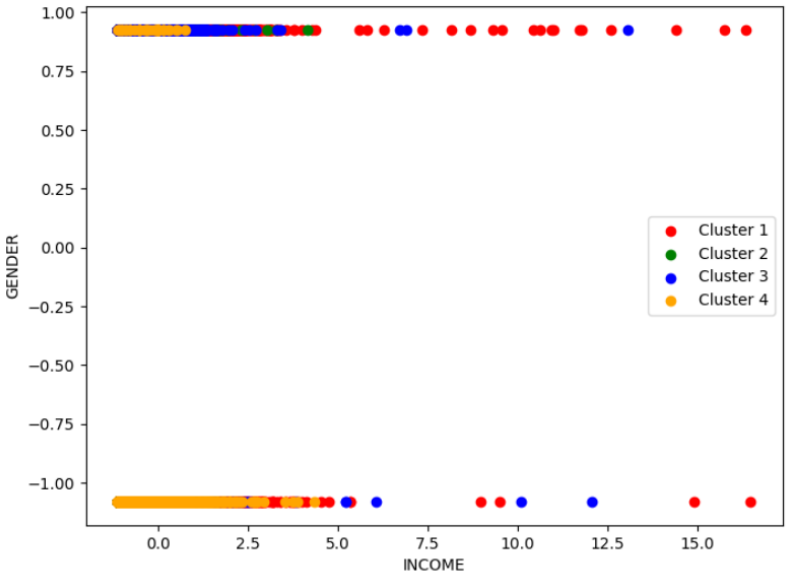
مقدار p-valueهای مربوط به تست‌های chi2:

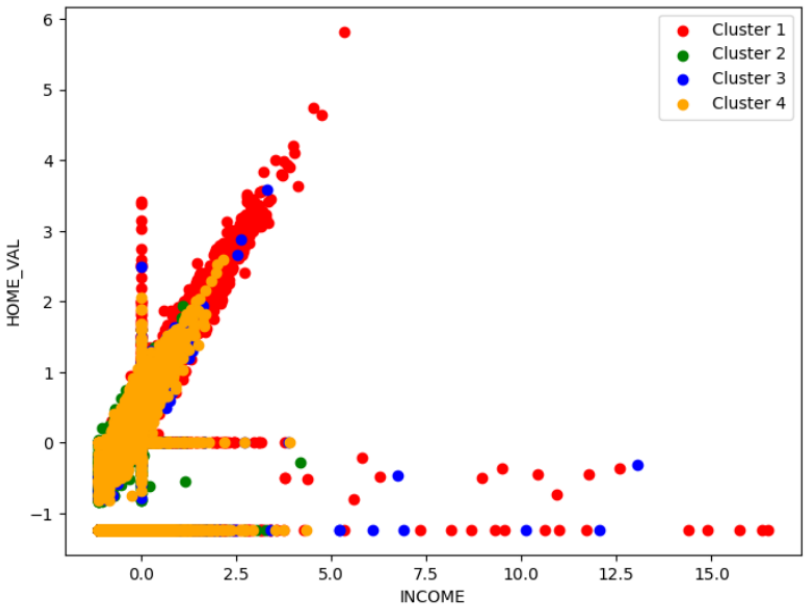
```
AGE : chi2 = 2109.8103858870645 , p-value = 0.0
HOMEKIDS : chi2 = 3130.8059348154015 , p-value = 0.0
PARENT1 : chi2 = 10170.159794431456 , p-value = 0.0
MSTATUS : chi2 = 2439.0089685988837 , p-value = 0.0
GENDER : chi2 = 5875.649413441442 , p-value = 0.0
EDUCATION : chi2 = 2061.8834648178167 , p-value = 0.0
OCCUPATION : chi2 = 2554.506014908983 , p-value = 0.0
CAR_TYPE : chi2 = 4706.412385848573 , p-value = 0.0
RED_CAR : chi2 = 4728.210465914627 , p-value = 0.0
CAR_AGE : chi2 = 1664.2335454950064 , p-value = 5.050671465501565e-288
CAR_USE : chi2 = 842.9685901026273 , p-value = 2.0751834302409707e-182
CLAIM_FLAG : chi2 = 650.0217541530201 , p-value = 1.440920839086806e-140
KIDSDRIV : chi2 = 640.4739440118128 , p-value = 2.3869255630812232e-129
YOJ : chi2 = 678.3433643952412 , p-value = 5.631217747918185e-104
CLM_FREQ : chi2 = 318.5669468753664 , p-value = 7.656288994447269e-59
URBANICITY : chi2 = 241.24544488531248 , p-value = 5.1188927465380735e-52
MVR_PTS : chi2 = 241.2656516176106 , p-value = 5.567086542450826e-31
BLUEBOOK : chi2 = 10331.537152279288 , p-value = 4.249985944291 38e-23
INCOME : chi2 = 26306.516371712016 , p-value = 1.401632569340874e-16
REVOKED : chi2 = 51.5123557509328 , p-value = 3.804765563660169e-11
CLM_AMT : chi2 = 7729.127965722419 , p-value = 6.921328927170983e-09
HOME_VAL : chi2 = 19430.0729591626 , p-value = 0.014518277600284051
TRAVTIME : chi2 = 348.796388705981 , p-value = 0.02068071544548019
TIF : chi2 = 87.55072836315625 , p-value = 0.039202173343841476
OLDCLAIM : chi2 = 10818.679063737436 , p-value = 0.10074208071702703
```

با استفاده از این اطلاعات، نمودار دو بعدی مربوط به چند مورد از ستون‌های با مقدار correlation بالا با لیبل خوشه‌ها (تاثیرگذار در خوشه بندی)، رسم شد:

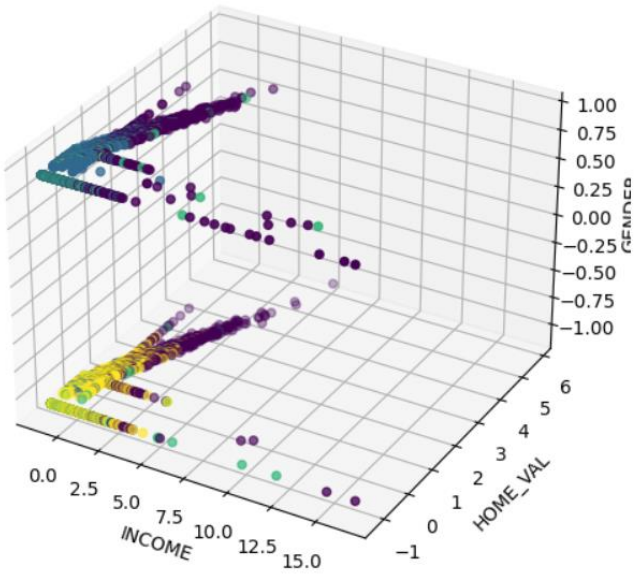

HOME_VAL
RED_CAR
Cluster 1
Cluster 2
Cluster 3
Cluster 4


GENDER
HOME_VAL
Cluster 1
Cluster 2
Cluster 3
Cluster 4


GENDER
INCOME
Cluster 1
Cluster 2
Cluster 3
Cluster 4




با توجه به نمایش دوبعدی خوشه بندی‌ها، برای نمایش ۳ بعدی آن‌ها، بهترین ۳ ستون برای نمایش آن‌ها، home-val، income و gender هستند:




بررسی خوشه‌ها این را نشان می‌دهد که یک خوشه شامل زنان دارای درامد پایین و خانه‌ی ارزان‌تر است. یک خوشه شامل مردان دارای درامد پایین و خانه‌ی ارزان‌تر است. یک خوشه شامل افراد دارای درامد بالا و خانه‌ی گران‌تر است. و خوشه‌ی آخر، شامل افراد دارای درامد بالا و خانه‌ی ارزان‌تر است.

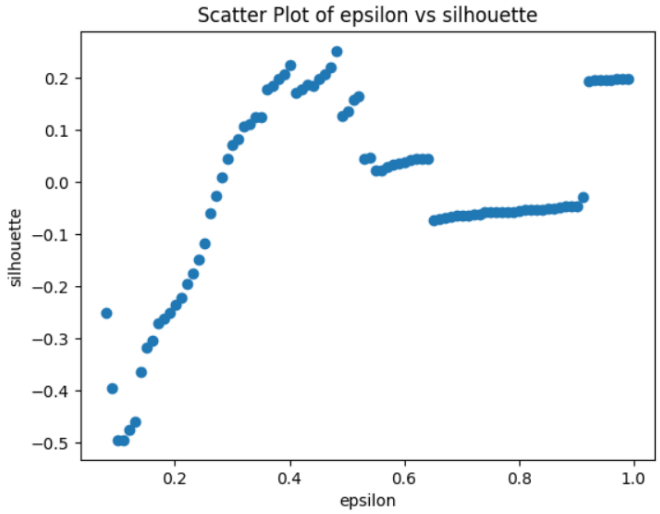
**سوال ۲:**

- **محاسبه‌ی مقدار بهینه‌ی epsilon و min-points با استفاده از معیار silhouette:**

مقدار epsilon در بازه‌ی ۰,۰۰۱ تا ۱ با فاصله‌های ۰,۰۱ تایی و مقدار min-points در بازه‌ی ۲ تا ۵۰ در نظر گرفته شد. برای هر یک از این حالت‌ها، مقدار silhouette محاسبه شد.

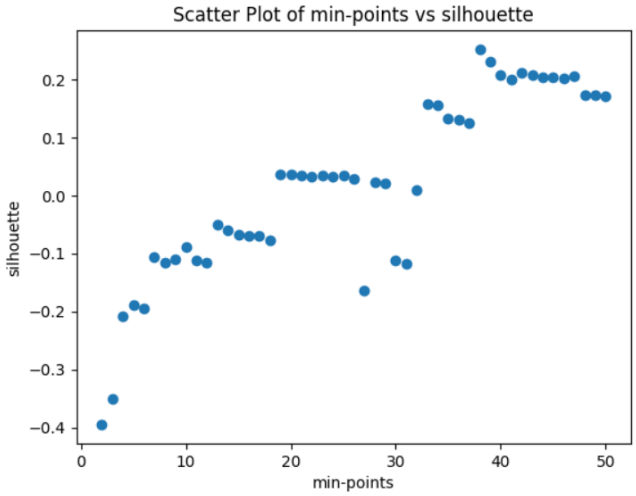
بهترین (بیشترین) مقدار برای silhouette، برابر با ۰,۲۵۲ (دقت برابر ۲۵ درصد) و در حالتی به دست آمد که epsilon=0.481 و min-points=38 قرار داده شد.

- تاثیر مقادیر مختلف epsilon در خوشه بندی:




همانطور که در نمودار هم نمایش داده شده است، در صورتی که مقدار epsilon خیلی کم باشد، دقت خوشه بندی بسیار پایین خواهد بود. اگر مقدار epsilon را تا حد متوسطی، زیاد کنیم، دقت مدل افزایش پیدا می‌کند. و در صورتی که مقدار آن را همچنان افزایش دهیم، دقت مدل به مرور کاهش می‌یابد. اما با افزایش epsilon تا جای ممکن، در انتهای بازه، دقت خوشه بندی افزایش می‌یابد. به طور تقریبی، بهترین دقت خوشه بندی، مربوط به زمانی‌ست که epsilon مقدار متوسطی داشته باشد.

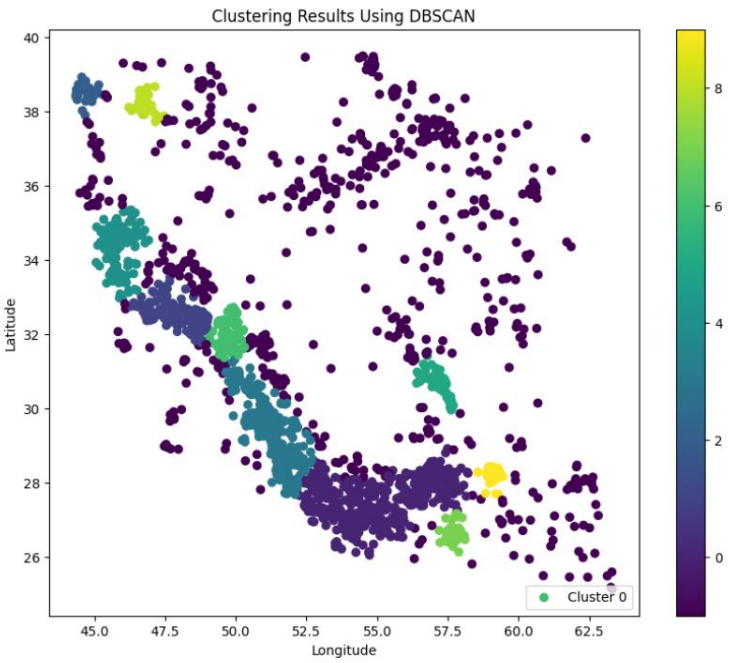
- تاثیر مقادیر مختلف min-points در خوشه بندی:




همانطور که در نمودار هم نمایش داده شده است، در صورتی که مقدار min-points خیلی کم باشد، دقت خوشه بندی بسیار پایین و در پایین‌تر حالت خواهد بود. با افزایش مقدار min-points، دقت مدل افزایش پیدا می‌کند. در صورتی که مقدار min-points خیلی زیاد باشد، دقت خوشه بندی به بیشترین حد خود می‌رسد.

- **اعمال الگوریتم DBSCAN با استفاده از مقادیر بهینه‌ی تعیین شده برای epsilon و min-points:**

خوشه‌های حاصل از این خوشه بندی، به شکل زیر حاصل شدند. (تعداد خوشه‌ها برابر ۱۰ است.):



- **افزودن ویژگی‌های بیشتری از مجموعه داده‌ها به مجموعه داده‌های انتخابی برای خوشه بندی:**

بقیه‌ی ویژگی‌های مجموعه داده‌ها به شرح زیر هستند:

```
['time',
 'latitude',
 'longitude',
 'depth',
 'mag',
 'magType',
 'nst',
 'gap',
 'dmin',
 'rms',
 'net',
 'id',
 'updated',
 'place',
 'type',
 'horizontalError',
 'depthError',
 'magError',
 'magNst',
 'status',
 'locationSource',
 'magSource']
```

ویژگی‌های زیر، از بین ویژگی‌های باقی مانده، برای خوشه بندی در نظر گرفته نشدند:

'time', 'magType', 'net', 'type', 'status', 'id', 'updated', 'locationSource', 'magSource' and 'place'

دلیل این کار این است که ویژگی‌ای مانند 'id'، تاثیری بر خوشه بندی نخواهد داشت. ویژگی‌ای مانند 'updated' که زمان به روز شدن اطلاعات مربوط به یک زلزله را نشان می‌دهد نیز، تاثیری بر خوشه بندی نخواهد داشت. و همچنین ویژگی‌ای مانند 'status'، با توجه به این‌که برای همه‌ی سطرها یکسان است، تاثیری بر خوشه بندی نخواهد داشت. تمامی این ویژگی‌ها، حداقل به یکی از این

۳ دلیل، تاثیری بر خوشه بندی ندارند و در محاسبات بعدی، از در نظر گرفتن آن‌ها صرف نظر شده‌است.

بقیه‌ی ویژگی‌های موجود در مجموعه داده‌ها و در نظر گرفته شده برای خوشه بندی، همگی عددی هستند و نیازی به تبدیل آن‌ها به مقادیر عددی نیست. تعدادی از آن‌ها دارای مقادیر گمشده هستند. با توجه به این‌که تعداد مقادیر گمشده‌ی آن‌ها زیاد است، قابل حذف نیستند. در نتیجه، مقادیر گمشده‌ی هر ستون، با میانگین آن ستون جایگزین شد.

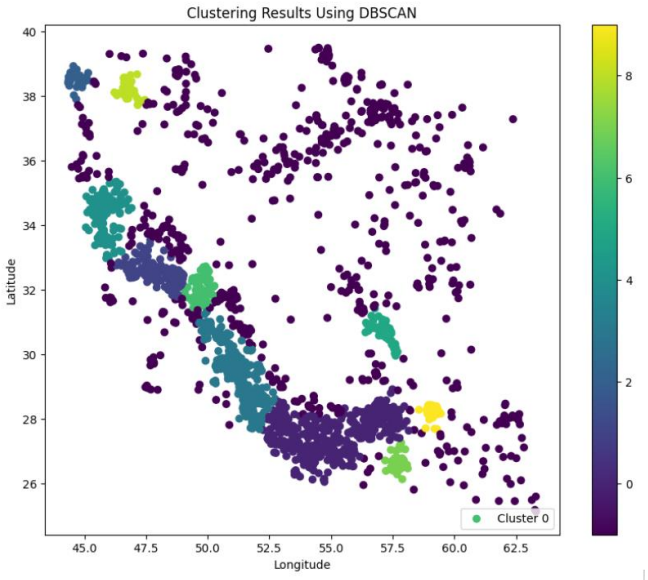
در نهایت، از بین زیرمجموعه‌های مجموعه داده‌ی مورد نظر، آن زیرمجموعه‌هایی که شامل هر دو ویژگی latitude و longitude بودند، در نظر گرفته شدند و هر بار با استفاده از یکی از آن‌ها، خوشه بندی DBSCAN با مقادیر بهینه‌ی بدست آمده، انجام شد و معیار silhouette برای سنجش دقت آن‌ها محاسبه شد.

بیشترین مقدار silhouette، برابر با ۰,۲۴۷ و در حالتی بدست آمد که مجموعه داده‌ی زیر در نظر گرفته شد:

[latitude, longitude, magError]

که این مقدار، همچنان از silhouette بدست آمده در زمانی که فقط دو ویژگی latitude و longitude برای خوشه بندی در نظر گرفته شده بودند، کمتر است.

خوشه‌های حاصل از این خوشه بندی، به شکل زیر حاصل شدند. (تعداد خوشه‌ها برابر ۱۰ است.):



پس، افزودن ویژگی‌های بیشتر به مجموعه داده‌های انتخابی برای خوشه بندی، موجب افزایش دقت خوشه بندی نمی‌شود؛ بلکه دقت آن را کاهش می‌دهد. و بهتر است از همان حالتی که فقط دو ویژگی latitude و longitude در نظر گرفته شده بودند، برای خوشه بندی استفاده کرد.

---

## ❖ سوالات تئوری:

### سوال ۱:

الگوریتم خوشه بندی Agglomerative، یک الگوریتم خوشه بندی پایین به بالاست، که در ابتدا هر یک از داده‌ها را یک خوشه در نظر می‌گیرد و هر بار هر یک از خوشه‌ها را با نزدیک‌ترین خوشه‌ی موجود، ترکیب می‌کند. آن‌قدر این کار را تکرار می‌کند تا همه‌ی داده‌ها در یک خوشه قرار بگیرند.

در ابتدا فاصله‌ی اقلیدسی هر یک از جفت داده‌ها محاسبه می‌شود:

$$
\begin{aligned}
|\mathrm{AB}| &= \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2} \\
&= \sqrt{(0.18 - 0.02)^2 + (0.76 - 0.27)^2} \\
&= \sqrt{(0.16)^2 + (0.49)^2} = \sqrt{0.025 + 0.241} \\
&= \sqrt{0.267} = 0.517
\end{aligned}
$$

$$
\begin{aligned}
|\mathrm{AC}| &= \sqrt{(x_A - x_C)^2 + (y_A - y_C)^2} \\
&= \sqrt{(0.18 - 0.55)^2 + (0.76 - 0.52)^2} \\
&= \sqrt{(-0.37)^2 + (0.24)^2} = \sqrt{0.137 + 0.058} \\
&= \sqrt{0.195} = 0.442
\end{aligned}
$$

$$|\mathrm{AD}| = \sqrt{(x_A - x_D)^2 + (y_A - y_D)^2}$$
$$= \sqrt{(0.18 - 0.88)^2 + (0.76 - 0.53)^2}$$
$$= \sqrt{(-0.7)^2 + (0.23)^2} = \sqrt{0.49 + 0.053} = \sqrt{0.543}$$
$$= 0.737$$

$$|\mathrm{AE}| = \sqrt{(x_A - x_E)^2 + (y_A - y_E)^2}$$
$$= \sqrt{(0.18 - 0.38)^2 + (0.76 - 0.77)^2}$$
$$= \sqrt{(-0.2)^2 + (-0.01)^2} = \sqrt{0.04 + 0} = \sqrt{0.04}$$
$$= 0.2$$

$$|\mathrm{AF}| = \sqrt{(x_A - x_F)^2 + (y_A - y_F)^2}$$
$$= \sqrt{(0.18 - 0.35)^2 + (0.76 - 0.05)^2}$$
$$= \sqrt{(-0.17)^2 + (0.71)^2} = \sqrt{0.029 + 0.504}$$
$$= \sqrt{0.533} = 0.73$$

$$|\mathrm{BC}| = \sqrt{(x_B - x_C)^2 + (y_B - y_C)^2}$$
$$= \sqrt{(0.02 - 0.55)^2 + (0.27 - 0.52)^2}$$
$$= \sqrt{(-0.53)^2 + (-0.25)^2} = \sqrt{0.28 + 0.063}$$
$$= \sqrt{0.343} = 0.586$$

$$|\mathrm{BD}| = \sqrt{(x_B - x_D)^2 + (y_B - y_D)^2}$$
$$= \sqrt{(0.02 - 0.88)^2 + (0.27 - 0.53)^2}$$
$$= \sqrt{(-0.86)^2 + (-0.26)^2} = \sqrt{0.74 + 0.068}$$
$$= \sqrt{0.808} = 0.899$$

$$|BE| = \sqrt{(x_B - x_E)^2 + (y_B - y_E)^2}$$
$$= \sqrt{(0.02 - 0.38)^2 + (0.27 - 0.77)^2}$$
$$= \sqrt{(-0.36)^2 + (-0.5)^2} = \sqrt{0.13 + 0.25} = \sqrt{0.38}$$
$$= 0.616$$

$$|BF| = \sqrt{(x_B - x_F)^2 + (y_B - y_F)^2}$$
$$= \sqrt{(0.02 - 0.35)^2 + (0.27 - 0.05)^2}$$
$$= \sqrt{(-0.33)^2 + (0.22)^2} = \sqrt{0.109 + 0.048}$$
$$= \sqrt{0.157} = 0.396$$

$$|CD| = \sqrt{(x_C - x_D)^2 + (y_C - y_D)^2}$$
$$= \sqrt{(0.55 - 0.88)^2 + (0.52 - 0.53)^2}$$
$$= \sqrt{(-0.33)^2 + (-0.01)^2} = \sqrt{0.109 + 0} = \sqrt{0.109}$$
$$= 0.33$$

$$|CE| = \sqrt{(x_C - x_E)^2 + (y_C - y_E)^2}$$
$$= \sqrt{(0.55 - 0.38)^2 + (0.52 - 0.77)^2}$$
$$= \sqrt{(0.17)^2 + (-0.25)^2} = \sqrt{0.029 + 0.063}$$
$$= \sqrt{0.092} = 0.303$$

$$|CF| = \sqrt{(x_C - x_F)^2 + (y_C - y_F)^2}$$
$$= \sqrt{(0.55 - 0.35)^2 + (0.52 - 0.05)^2}$$
$$= \sqrt{(0.2)^2 + (0.47)^2} = \sqrt{0.04 + 0.221} = \sqrt{0.261}$$
$$= 0.511$$

$$|\mathrm{DE}| = \sqrt{(x_D - x_E)^2 + (y_D - y_E)^2}$$
$$= \sqrt{(0.88 - 0.38)^2 + (0.53 - 0.77)^2}$$
$$= \sqrt{(0.5)^2 + (-0.24)^2} = \sqrt{0.25 + 0.058} = \sqrt{0.308}$$
$$= 0.555$$

$$|\mathrm{DF}| = \sqrt{(x_D - x_F)^2 + (y_D - y_F)^2}$$
$$= \sqrt{(0.88 - 0.35)^2 + (0.53 - 0.05)^2}$$
$$= \sqrt{(0.53)^2 + (0.48)^2} = \sqrt{0.281 + 0.23} = \sqrt{0.511}$$
$$= 0.715$$

$$|\mathrm{EF}| = \sqrt{(x_E - x_F)^2 + (y_E - y_F)^2}$$
$$= \sqrt{(0.38 - 0.35)^2 + (0.77 - 0.05)^2}$$
$$= \sqrt{(0.03)^2 + (0.72)^2} = \sqrt{0 + 0.519} = \sqrt{0.519}$$
$$= 0.72$$

ابتدا هر یک از داده‌ها را به عنوان یک خوشه‌ی جدا در نظر می‌گیریم:

[A], [B], [C], [D], [E], [F]

- Single link: کم‌ترین فاصله‌ی بین نقاط یک خوشه با خوشه‌ی دیگر، به عنوان فاصله‌ی خوشه‌ها در نظر گرفته می‌شود.

در ابتدا خوشه‌های مربوط به دو داده‌ی A و E با هم ترکیب می‌شوند:

[A, E], [B], [C], [D], [F]

سپس، دو خوشه‌ی [A, E] و [C] با هم ترکیب می‌شوند:

[A, E, C], [B], [D], [F]

سپس، دو خوشه‌ی [A, E, C] و [D] با هم ترکیب می‌شوند:

[A, E, C, D], [B], [F]

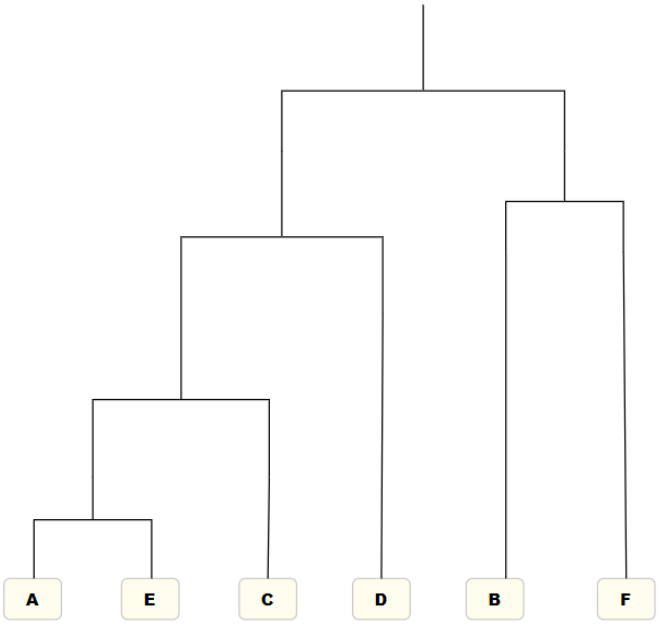
سپس، دو خوشه‌ی [B] و [F] با هم ترکیب می‌شوند:

[A, E, C, D], [B, F]

و در نهایت، دو خوشه‌ی [A, E, C, D] و [B, F] با هم ترکیب می‌شوند:

[A, E, C, D, B, F]

نمودار dendrogram مربوطه:




- complete link: بیشترین فاصله‌ی بین نقاط یک خوشه با خوشه‌ی دیگر، به عنوان فاصله‌ی خوشه‌ها در نظر گرفته می‌شود.

در ابتدا خوشه‌های مربوط به دو داده‌ی A و E با هم ترکیب می‌شوند:

[A, E], [B], [C], [D], [F]

سپس، دو خوشه‌ی [C] و [D] با هم ترکیب می‌شوند:

[A, E], [B], [C, D], [F]

سپس، دو خوشه‌ی [B] و [F] با هم ترکیب می‌شوند:

[A, E], [C, D], [B, F]

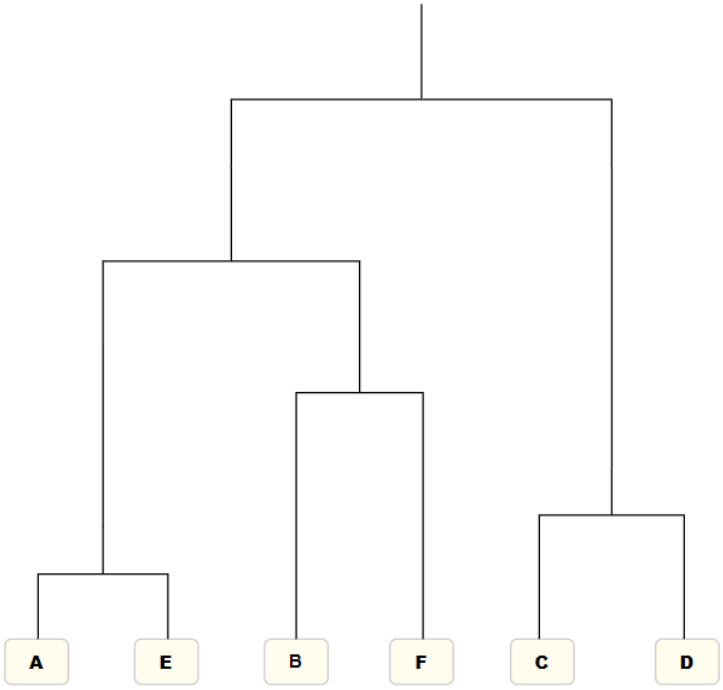
سپس، دو خوشه‌ی [A, E] و [B, F] با هم ترکیب می‌شوند:

[A, E, B, F], [C, D]

و درنهایت، دو خوشه‌ی [A, E, B, F] و [C, D] با هم ترکیب می‌شوند:

[A, E, B, F, C, D]

نمودار dendrogram مربوطه:



**سوال ۲:**

- **اعمال الگوریتم خوشه بندی k-modes:**

ابتدا باید فاصله‌ی نقاط، با هر یک از centroidها محاسبه شود، تا تعیین شود که هر یک از نقاط، در خوشه‌ی مربوط به کدام یک از centroidها قرار می‌گیرد:

| | Cluster1 (3) | Cluster2 (8) | Cluster3 (9) | Cluster |
|---|---|---|---|---|
| 0 | 5 | 5 | 2 ✓ | 2 |
| 1 | 3 ✓ | 4 | 3 | 1 |
| 2 | 3 ✓ | 3 | 3 | 1 |
| 3 | 0 ✓ | 4 | 4 | 1 |
| 4 | 3 | 4 | 2 ✓ | 3 |
| 5 | 2 ✓ | 4 | 4 | 1 |
| 6 | 4 | 2 ✓ | 5 | 2 |
| 7 | 4 | 3 ✓ | 3 | 2 |
| 8 | 4 | 0 ✓ | 5 | 2 |
| 9 | 4 | 5 | 0 ✓ | 3 |

با توجه به فاصله‌ها، خوشه‌ها بدین شرح حاصل می‌شوند:

Cluster 1 → 1, 2, 3, 5
Cluster 2 → 0, 6, 7, 8
Cluster 3 → 4, 9

حال برای هر یک از خوشه‌ها، mode آن‌ها محاسبه می‌شود و به عنوان centroid جدید در نظر گرفته می‌شود تا اعضای خوشه‌ها به روز شود:

Cluster 1 → [Extroverted, Direct, Emotional, Analytical, Democratic]
Cluster 2 → [Introverted, Assertive, Intuitive, Creative, Authoritarian]
Cluster 3 → [Extroverted, Assertive, Logical, Analytical, Authoritarian]

حال با توجه به این centroidهای جدید، فاصله‌ی نقطه‌ها تا خوشه‌ها و در نتیجه خوشه‌ها به روز می‌شوند:

| | Cluster1 | Cluster2 | Cluster3 | Cluster |
|---|---|---|---|---|
| 0 | 4 | 3 ✓ | 3 | 2 |
| 1 | 2 ✓ | 3 | 2 | 1 |
| 2 | 2 ✓ | 3 | 3 | 1 |
| 3 | 1 ✓ | 5 | 4 | 1 |
| 4 | 4 | 3 | 1 ✓ | 3 |
| 5 | 3 ✓ | 4 | 4 | 1 |
| 6 | 5 | 2 ✓ | 4 | 2 |
| 7 | 4 | 1 ✓ | 2 | 2 |
| 8 | 4 | 2 ✓ | 4 | 2 |
| 9 | 3 | 4 | 1 ✓ | 3 |

با توجه به فاصله‌ها، خوشه‌ها بدین شرح حاصل می‌شوند:

Cluster 1 → 1, 2, 3, 5
Cluster 2 → 0, 6, 7, 8
Cluster 3 → 4, 9

با توجه به این‌که در اعضای متعلق به خوشه‌ها تغییری حاصل نشد، الگوریتم در همین‌جا متوقف می‌شود. و خوشه‌های حاصل شده در این مرحله، نتیجه‌ی خوشه بندی با k-modes، خواهند بود:

[1, 2, 3, 5], [0, 6, 7, 8], [4, 9]

- محاسبه‌ی معیارهای **precision**، **recall**، **f1** و **entropy**:

$$\textbf{precision} = \frac{\text{points related to majority class in the cluster}}{\text{all of the points in the cluster}}$$

$$\text{precision}_{\text{cluster\_1}} = \frac{540}{2 + 22 + 540} = \frac{540}{564} = 0.957$$

$$\text{precision}_{\text{cluster\_2}} = \frac{333}{23 + 333 + 242 + 89} = \frac{333}{687} = 0.485$$

$$\text{precision}_{\text{cluster\_3}} = \frac{700}{36 + 12 + 700 + 1} = \frac{700}{749} = 0.935$$

$$\textbf{recall} = \frac{\text{points related to majority class in the cluster}}{\text{all of the points related to majority class}}$$

$$\text{recall}_{\text{cluster\_1}} = \frac{540}{540 + 89 + 1} = \frac{540}{630} = 0.857$$

$$\text{recall}_{\text{cluster\_2}} = \frac{333}{333 + 12} = \frac{333}{345} = 0.965$$

$$\text{recall}_{\text{cluster\_3}} = \frac{700}{22 + 242 + 700} = \frac{700}{964} = 0.726$$

$$\mathbf{f1} = 2 * \frac{\text{precision} * recall}{\text{precision} + recall}$$

$$\begin{aligned} \text{f1}_{\text{cluster\_1}} &= 2 * \frac{\text{precision}_{\text{cluster\_1}} * \text{recall}_{\text{cluster\_1}}}{\text{precision}_{\text{cluster\_1}} + \text{recall}_{\text{cluster\_1}}} \\ &= 2 * \frac{0.957 * 0.857}{0.957 + 0.857} = 2 * \frac{0.82}{1.814} = 2 * 0.452 \\ &= 0.904 \end{aligned}$$

$$\begin{aligned} \text{f1}_{\text{cluster\_2}} &= 2 * \frac{\text{precision}_{\text{cluster\_2}} * \text{recall}_{\text{cluster\_2}}}{\text{precision}_{\text{cluster\_2}} + \text{recall}_{\text{cluster\_2}}} \\ &= 2 * \frac{0.485 * 0.965}{0.485 + 0.965} = 2 * \frac{0.468}{1.45} = 2 * 0.323 \\ &= 0.646 \end{aligned}$$

$$\begin{aligned} \text{f1}_{\text{cluster\_3}} &= 2 * \frac{\text{precision}_{\text{cluster\_3}} * \text{recall}_{\text{cluster\_3}}}{\text{precision}_{\text{cluster\_3}} + \text{recall}_{\text{cluster\_3}}} \\ &= 2 * \frac{0.935 * 0.726}{0.935 + 0.726} = 2 * \frac{0.679}{1.661} = 2 * 0.409 \\ &= 0.818 \end{aligned}$$

$$\mathbf{Entropy} = \text{H(C)} = -\sum_{\text{i=1}}^{\text{r}} \text{p}_{\text{C}_\text{i}} \log_2 \text{p}_{\text{C}_\text{i}}$$

$$\begin{aligned} \text{Entropy}_{\text{cluster\_1}} &= \text{H(cluster\_1)} = -\sum_{\text{i=1}}^{4} \text{p}_{\text{C}_\text{i}} \log_2 \text{p}_{\text{C}_\text{i}} \\ &= -\left(\frac{2}{564}\log_2\frac{2}{564} + \frac{22}{564}\log_2\frac{22}{564} + \frac{540}{564}\log_2\frac{540}{564}\right) \\ &= -(0.004\log_2 0.004 + 0.039\log_2 0.039 \\ &+ 0.957\log_2 0.957) \\ &= -\big(0.004 * (-7.966) + 0.039 * (-4.68) + 0.957 \\ &* (-0.063)\big) = -(-0.032 - 0.183 - 0.06) = 0.275 \end{aligned}$$

$$\text{Entropy}_{\text{cluster\_2}} = \text{H}(\text{cluster\_2}) = -\sum_{i=1}^{4} p_{C_i} \log_2 p_{C_i}$$

$$= -\left(\frac{23}{687}\log_2\frac{23}{687} + \frac{333}{687}\log_2\frac{333}{687} + \frac{242}{687}\log_2\frac{242}{687} + \frac{89}{687}\log_2\frac{89}{687}\right)$$

$$= -(0.033\log_2 0.033 + 0.485\log_2 0.485 + 0.352\log_2 0.352 + 0.13\log_2 0.13)$$

$$= -\big(0.033 * (-4.921) + 0.485 * (-1.044) + 0.352 * (-1.506) + 0.13 * (-2.943)\big)$$

$$= -(-0.162 - 0.506 - 0.53 - 0.383) = 1.581$$

$$\text{Entropy}_{\text{cluster\_3}} = \text{H}(\text{cluster\_3}) = -\sum_{i=1}^{4} p_{C_i} \log_2 p_{C_i}$$

$$= -\left(\frac{36}{749}\log_2\frac{36}{749} + \frac{12}{749}\log_2\frac{12}{749} + \frac{700}{749}\log_2\frac{700}{749} + \frac{1}{749}\log_2\frac{1}{749}\right)$$

$$= -(0.048\log_2 0.048 + 0.016\log_2 0.016 + 0.935\log_2 0.935 + 0.001\log_2 0.001)$$

$$= -\big(0.048 * (-4.381) + 0.016 * (-5.966) + 0.935 * (-0.097) + 0.001 * (-9.966)\big)$$

$$= -(-0.021 - 0.095 - 0.091 - 0.01) = 0.217$$