به نام خداوند بخشنده و مهربان

درس داده کاوی

تمرین دوم

استاد درس: دکتر ناظرفرد

نام دانشجو:

روزبه قاسمی ۹۵۳۱۴۲۴

اردیبهشت ۱۳۹۹

سوال اول

ایده Mini-Batch KMeans یعنی ایده اصلی الگوریتم استفاده از دسته‌های تصادفی کوچک داده از یک اندازه ثابت است، بنابراین آن‌ها می‌توانند در حافظه ذخیره شوند. هر تکرار یک نمونه تصادفی جدید از مجموعه داده‌ها بدست می‌آید و برای به روز رسانی خوشه‌های مورد استفاده قرار می‌گیرد و این امر تا زمانی تکرار می‌شود که دسته‌ای کوچک با استفاده از ترکیب محدب از مقادیر اولیه و داده‌ها، با استفاده از یک نرخ یادگیری که با تعداد تکرارها کاهش می‌یابد. این نرخ یادگیری معکوس تعداد داده‌های اختصاص‌داده‌شده به یک خوشه در طول فرآیند است. همانطور که تعداد تکرارها افزایش می‌یابد، اثر داده‌های جدید کاهش می‌یابد، بنابراین هم‌گرایی را می توان تشخیص داد وقتی هیچ تغییری در خوشه‌ها در چندین تکرار متوالی رخ نمی‌دهد. نتایج تجربی نشان می‌دهد که می‌تواند یک صرفه‌جویی قابل‌توجه در زمان محاسباتی به هزینه برخی از از دست دادن کیفیت خوشه بدست آورد، اما مطالعه گسترده این الگوریتم برای اندازه‌گیری این که چگونه ویژگی‌های مجموعه داده‌ها، مانند تعداد خوشه‌ها یا اندازه آن، بر کیفیت افراز تاثیر می‌گذارد، انجام شده‌است. این الگوریتم به طور تصادفی مجموعه‌ای از مجموعه داده‌ها را برای هر تکرار انتخاب می‌کند.
حال به معایب آن می‌پردازیم. از آنجایی که این روش مشابه روش k-means است پس معایب آن نیز مشابه همان است:

۱- سختی در بدست آوردن K مناسب.
۲- قسمت‌های اولیه مختلف می‌توانند منجر به خوشه‌های نهایی متفاوتی شوند.
۳- با افزایش تعداد تکرارها، کیفیت خوشه‌بندی کاهش می‌یابد و داده‌های جدید بر خوشه‌ها و مراکز جرم تاثیر کمتری خواهند داشت.
۴- این روش نسبت به داده‌های دورافتاده و نویز حساس است، اما در این مورد بهتر از الگوریتم k-means عمل می‌کند.

سوال دوم

۱) استفاده از RBF

شبکه عصبی RBF یک شبکه عصبی Feed-Forward با قابلیت تقریب قوی است. یک الگوریتم k-means براساس پارامتر چگالی برای تعیین مرکز خوشه‌بندی با هدف بهبود نرخ آموزش RBF در این مقاله استفاده شد. به علت طولانی بودن مطالب ذکر شده در مقاله به صورت خلاصه آنرا ذکر میکنم. الگوریتم k-means به عنوان الگوریتم یادگیری در شبکه عصبی RBF به کار می‌رود. مراکز اولیه خوشه‌بندی به طور تصادفی براساس الگوریتم k-means معمولی انتخاب شده‌اند که منجر به توانایی‌های هم‌گرایی متفاوتی می‌شوند. در این مقاله روش پارامترهای چگالی برای کاهش حساسیت انتخاب تصادفی مراکز خوشه‌ها در این مقاله و محاسبه پهنای تابع پایه شبکه RBF مورد استفاده قرار گرفته‌است. مقایسه با الگوریتم‌های سنتی، k-means مدل شبکه RBF براساس پارامتر چگالی از مزایای آن در تعیین تعداد نتایج خوشه‌بندی و مراکز توابع پایه برخوردار است و مدل دارای نرخ هم‌گرایی بالاتر و دقت پیش‌بینی بود.

۲) **استفاده از روش MacQueen**

این الگوریتم با هدف جداسازی n اشیاء در گروه های غیر همپوشانی به عنوان حداقل خطاهای مربع (یعنی مجموع فاصله بین نقاط و مرکز گروه آنها) به حداقل می رسد. در مرحله اول ، این نوع از k-means به مرحله ای از انتخاب اولیه می رود و نقاط نقاط داده k را به عنوان سنترویدها (مراکز پارتیشن) انتخاب می کند و نقاط را به نزدیکترین سانتروئید با توجه به فاصله اقلیدسی و به روز کردن سنترویدها با استفاده از میانگین نقاط در گروه سپس، الگوریتم بصورت تکراری تا زمانی که همگرایی به یک تکلیف مرحله ای انجام شود که در آن هر نقطه با توجه به فاصله اقلیدسی به نزدیکترین سنترویدِ اختصاص یابد و سنترویدِ مربوطه با استفاده از میانگین نقاط در گروه به روز شود. همگرایی هنگامی حاصل می شود که سنترویدها از حرکت خود متوقف می شوند یا وقتی تعداد تکرارهای داخلی به دست می آید. کیفیت خوشه بندی تولید شده توسط الگوریتم MacQueen k-mean توسط شاخص اعتبار خوشه کالینسکی-هاراباسز مشهور ارزیابی می شود (کاللیسکی و هاراباسز ، ۱۹۷۴).

مراحل مربوط به این الگوریتم شامل:

۱) K شئ به طور تصادفی شده و از آنها به عنوان سنترویدِ اولیه استفاده می‌شود.
۲) هر داده را با نزدیک‌ترین سنترویدِ مشخص کنید.
۳) پس از انجام هربار این اعمال، مراکز سنترویدها می‌بایست بروز شود.

سوال سوم

الف) درست است به این دلیل که شعاع تعریف‌شده به صورت اپسیلون فاصله از نقطه مرکزی است که تمام همسایه‌های آن باید از نقطه مرکزی کم‌تر یا مساوی باشند. دو نقطه نیز همسایه هستند که از طریق زنجیره‌ای از نقاط که دو یا دو همسایه هستند قابل‌دسترس هستند، بنابراین نقاط کانونی بسیاری در یک خوشه وجود دارد و هر نقطه در خوشه باید در آستانه فاصله تا نقطه مرکزی باشد.

ب) اشتباه است به این دلیل که الگوریتم dbcsan، یک خوشه از هر شکل دل‌خواه را تشکیل می‌دهد و فرضیات قوی برای توزیع نقاط داده در تصویر ندارد.

پ) اشتباه است به این دلیل که الگوریتم dbscan دارای پیچیدگی زمانی پایین O(nlogn) است.

ت) درست است به این دلیل که این الگوریتم نقاط داده را براساس چگالی داده خوشه‌بندی می‌کند و نیازی به دانستن تعداد خوشه‌ها ندارد.

ث) درست است به این دلیل که الگوریتم dbscan برای داده‌های دورافتاده robust است. روش dbscan برای جداسازی خوشه‌های با تراکم زیاد از خوشه‌های کم‌تراکم بسیار عالی است . روش dbscan بر خلاف روش k-means نیازمند تعداد خوشه‌های مشخص نیست.

سوال چهارم

الف) Single link

	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.84	0.16	0.14	0		
E	0.28	0.77	0.70	0.45	0	
F	0.34	0.61	0.93	0.20	0.67	0

Distance metric 0

	A,B	C	D	E	F
A,B	0				
C	0.25	0			
D	0.16	0.14	0		
E	0.28	0.70	0.45	0	
F	0.34	0.93	0.20	0.67	0

Distance metric 1

	A,B	C,D	E	F
A,B	0			
C,D	0.16	0		
E	0.28	0.45	0	
F	0.34	0.20	0.67	0

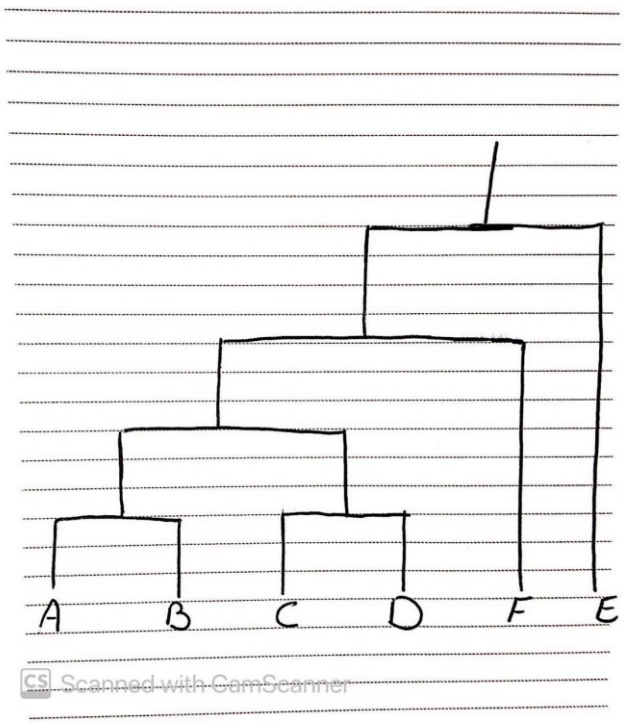
Distance metric 2

	A,B,C,D	E	F
A,B,C,D	0		
E	0.28	0	
F	0.20	0.67	0

Distance metric 3

	A,B,C,D,F	E
A,B,C,D,F	0	
E	0.28	0

Distance metric 4



Complete link ب

	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.84	0.16	0.14	0		
E	0.28	0.77	0.70	0.45	0	
F	0.34	0.61	0.93	0.20	0.67	0

Distance metric 0

	A,B	C	D	E	F
A,B	0				
C	0.51	0			
D	0.84	0.14	0		
E	0.77	0.70	0.45	0	
F	0.61	0.93	0.20	0.67	0

Distance metric 1

	A,B	C,D	E	F
A,B	0			
C,D	0.84	0		
E	0.77	0.70	0	
F	0.61	0.93	0.67	0

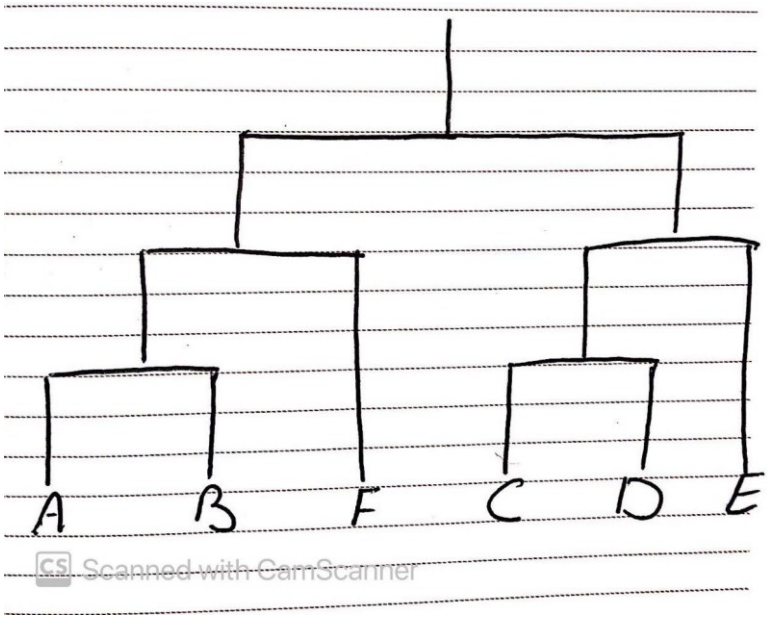
Distance metric 2

	A,B,F	C,D	E
A,B,F	0		
C,D	0.93	0	
E	0.77	0.70	0

Distance metric 3

	A,B,F	C,D,E
A,B,F	0	
C,D,E	0.93	0

Distance metric 4



سوال پنجم

الف) فقط با روش dbscan. این یک مجموعه داده محدب است و می‌تواند به‌صورت خطی تقسیم‌بندی شود که روش الگوریتم-k means است. روش dbscan را می‌توان با استفاده از الگوریتم مبتنی‌بر چگالی خوشه‌بندی کرد.

ب) هر دو روش dbscan و k-means. در این مجموعه داده‌ها اگر تعداد خوشه‌ها را به الگوریتم k-means بدهیم، این مجموعه داده‌ها را به وضوح مانند تصویر نشان‌داده‌شده در بالا خوشه‌بندی خواهد کرد زیرا خوشه‌ها می‌توانند به‌صورت خطی تقسیم‌بندی شوند.

پ) تنها روش الگوریتم k-means درست است. اگر تعداد خوشه‌ها را به الگوریتم k-means بدهیم، داده‌ها را به همان اندازه که در بالا نشان‌داده شده‌است، خوشه‌بندی می‌کند. با این حال در حالی که چگالی داده تقریباً در تمام فضای داده‌ها یک‌سان است، این مجموعه داده‌ها را به صورت یک خوشه در نظر می‌گیرد یا اگر پارامترهای الگوریتم dbscan را انتخاب کرده و اپسیلون کمتری انتخاب کنید، داده‌های نویز زیادی را شناسایی خواهد کرد.

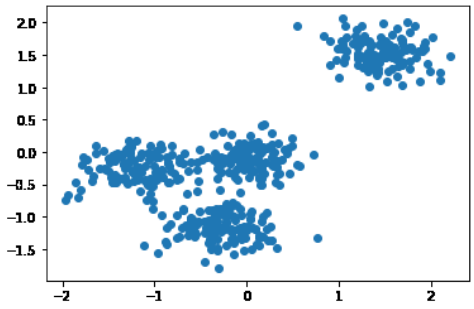
ت) فقط با روش dbscan . مشابه شکل الف، این مجموعه داده دل‌خواه است و می‌تواند توسط الگوریتم k-means تقسیم شود در حالی که داده‌های قابل‌تفکیک خطی نیستند. با استفاده از ویژگی همبندی می‌توان دو خوشه را تشخیص داد.

گزارش بخش پیاده‌سازی

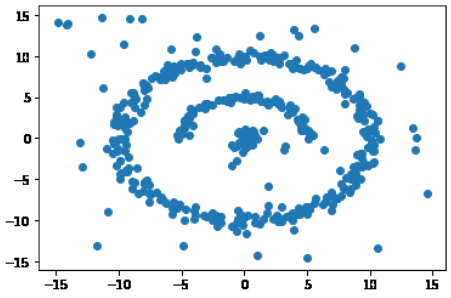
پیاده سازی اول ؛ الگوریتم K-means

در این بخش از پیاده‌سازی‌ها، الگوریتم k-means با زبان برنامه‌نویسی پایتون را برای خوشه‌بندی مجموعه داده‌های دو بعدی پیاده‌سازی می‌شود. در تمامی پیاده سازی ها به دلیل سهولت استفاده از گوگل کولب، کدها همگی در گوگل کولب پیاده‌سازی شده است.

نمایش دیتاست های اول و دوم به ترتیب شکل‌های زیر هستند:



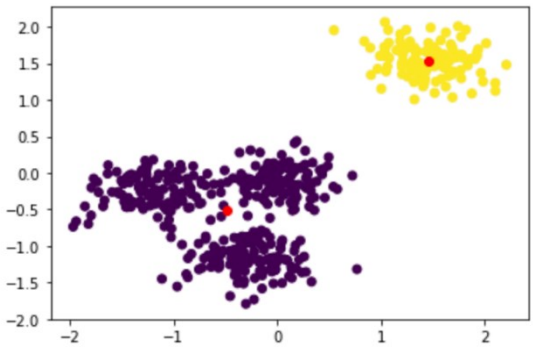
دیتاست اول



دیتاست دوم

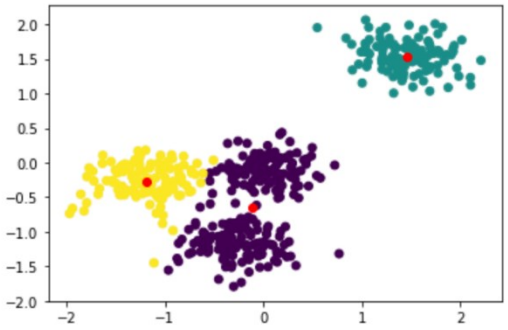
قسمت A :

نتایج این قسمت به شکل‌های زیر است:



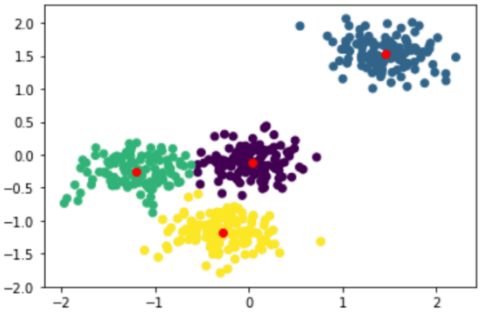
```
Cluster errors:  [0.74911603 0.31600882]
```

برای k=2 و به تعداد iteration = 20



```
Cluster errors:  [0.61650061 0.31600882 0.33740166]
```

برای k=3 و به تعداد iteration = 20



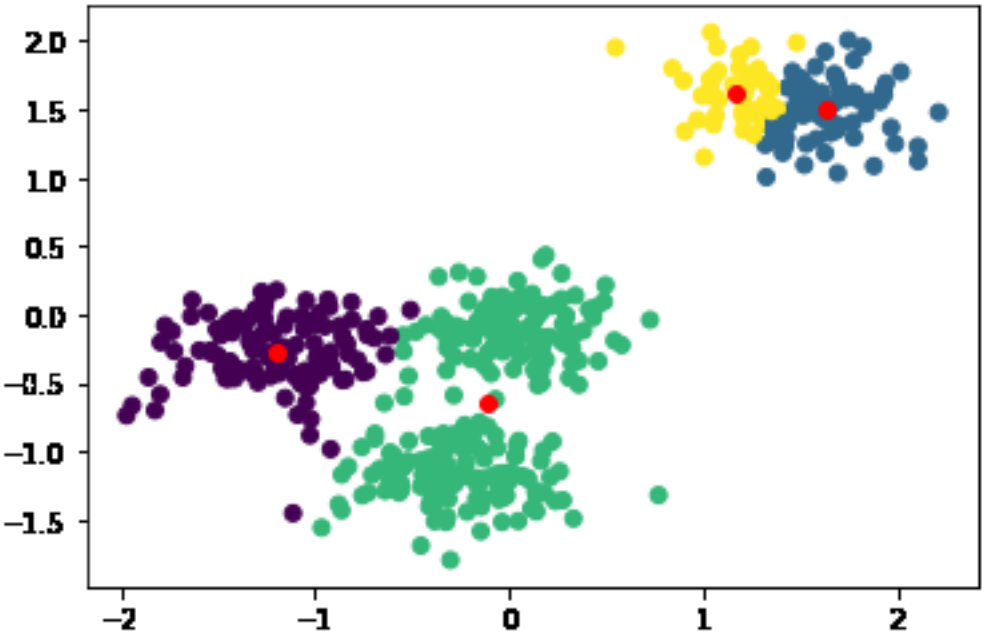
```
Cluster errors:  [0.28865974 0.31600882 0.32383689 0.33612332]
```

برای k=4 و به تعداد iteration = 20

قسمت B : میانگین فواصل خوشه‌ها در تصاویر بالا به عنوان خطای خوشه چاپ می‌شود.

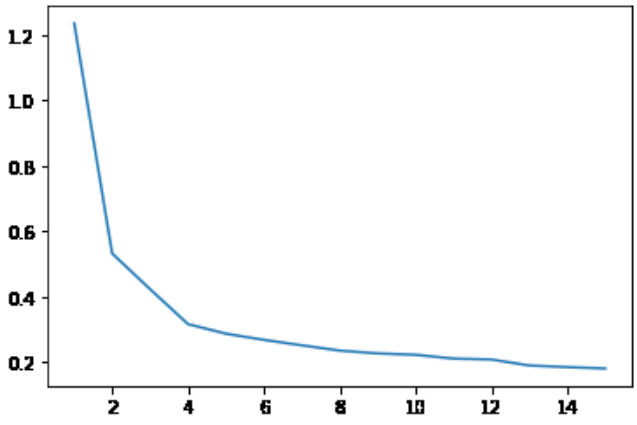
قسمت C :

میانگین خطاهای خوشه‌بندی با استفاده از روش k-means بدست‌آمده است. اگر به حداقل برسانیم، به راه‌حل خوشه‌بندی کامل رسیده‌ایم. در تصویر زیر، میانگین خطای خوشه‌بندی k = 6 روی همان مجموعه داده را مشاهده می‌کنیم:



Cluster errors: [0.33740166 0.25584071 0.61449976 0.2295376]

قسمت D : در این بخش خطای خوشه‌بندی k = 1 تا k = 15 را در گراف رسم شده مشاهده می‌کنیم:

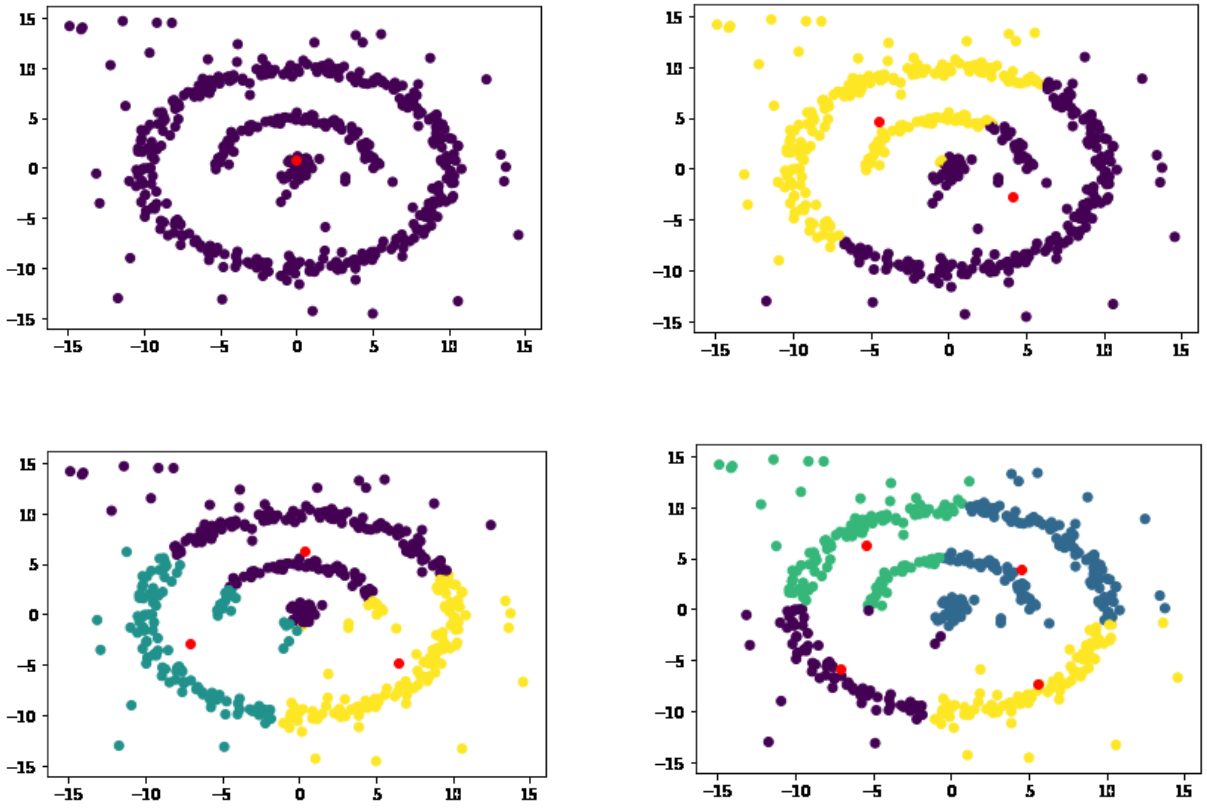


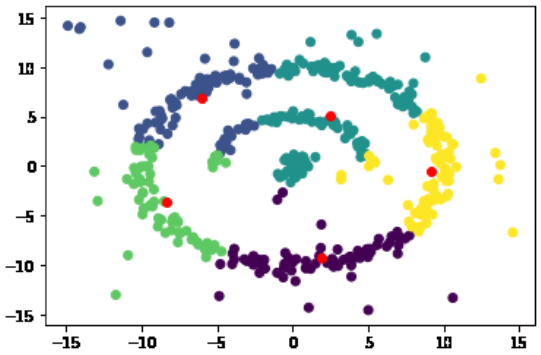
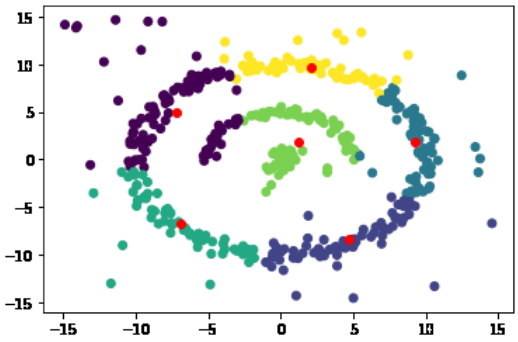
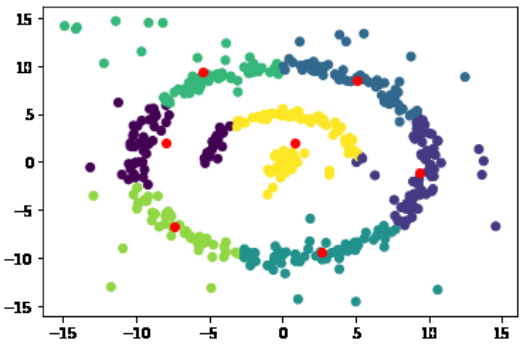
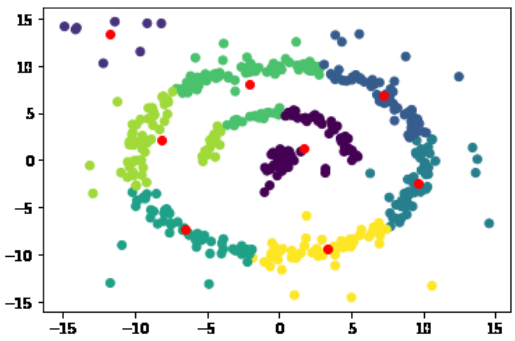
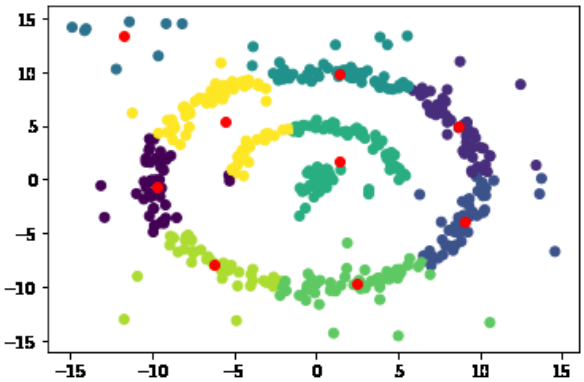
قسمت E :

الگوریتم elbow به ما می گوید که در نمودار WCSS_K جائی وجود دارد و مانند شکل آرنج است و این بدان معنی است که هیچ تغییر بزرگ دیگری در کل خوشه‌ای از اندازه مربع وجود ندارد و ما برای روش خوشه بندی خود به K مناسب می‌رسیم. ما می‌توانیم $k = 4$ یا $k = 5$ را به عنوان تعداد مناسب خوشه ها برای این مجموعه داده انتخاب کنیم.

قسمت F :

مانطور که در تصویر داده خام Dataset2 مشاهده می کنید (قبل از خوشه بندی) یک شکل محدب وجود دارد که نمی تواند با استفاده از خطوط تقسیم شود، این بدان معناست که تقسیم خطی نیست بنابراین الگوریتم k-means نمی تواند این داده ها را در خوشه های مناسب تقسیم کنید. در این مجموعه داده ها ، طبقه بندی k به این معنی که k ای وجود دارد که در زیر با k=2 یا k=3 یا k=4 نشان داده شده است:

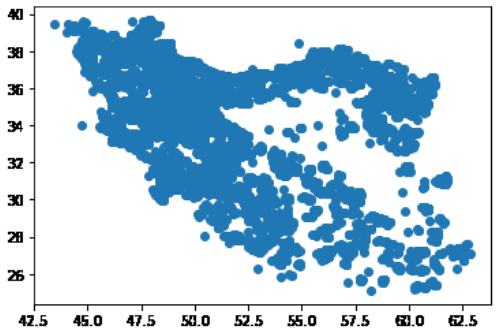


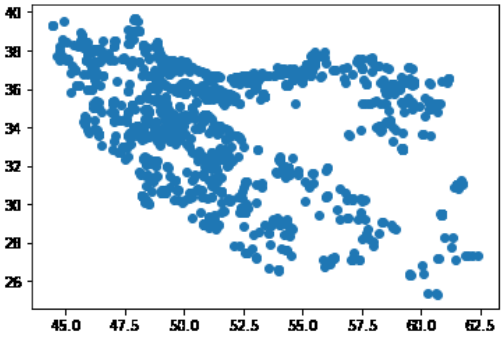






شکل نهایی بدست آمده

پیاده سازی دوم ؛ الگوریتم dbscan

در این بخش می‌خواهیم از الگوریتم dbscan (خوشه‌بندی فضایی مبتنی‌بر چگالی) برای تشخیص خوشه‌های متراکم بیماری در ایران استفاده کنیم . الگوریتم dbscan، الگوریتم خوشه‌بندی مبتنی‌بر چگالی است که می‌تواند شکل محدب داده‌ها را به خوبی خوشه‌بندی کرده و نواحی متراکم را براساس دو پارامتر نشان دهد : Minpts که نماینده حداقل تعداد نقاط است که یک نقطه مرکزی باید در شعاع مشخص‌شده توسط ۲- اپسیلون باشد. همانند بخش قبلی روش dbscan را برای الگوریتم dbscan و کتابخانه sklearn.cluster که به ما اجازه می‌دهد تا نقشه دنیای واقعی را بارگذاری کنیم و نقاط داده را روی نقشه ترسیم کنیم. ابتدا فایل csv می‌خوانیم و آن را در ردیف قبلی نگهدارید. سپس نقاط داده را با نمودار پراکندگی ساده رسم می‌کنیم که در شکل زیر نشان‌داده شده‌است .



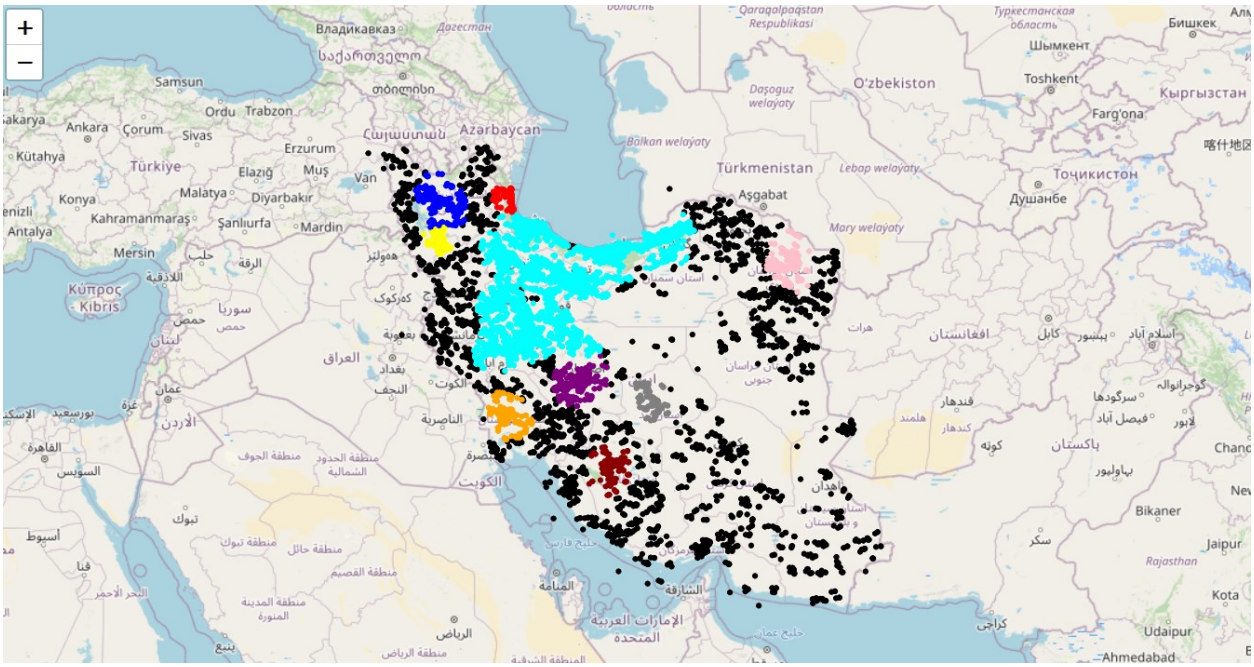


در این خط کد از فولیوم برای ترسیم نقشه واقعی با محل شروع در ایران و بزرگنمایی ۵ در ابتدا استفاده می کنیم. نتیجه را می توانید در شکل زیر مشاهده کنید:

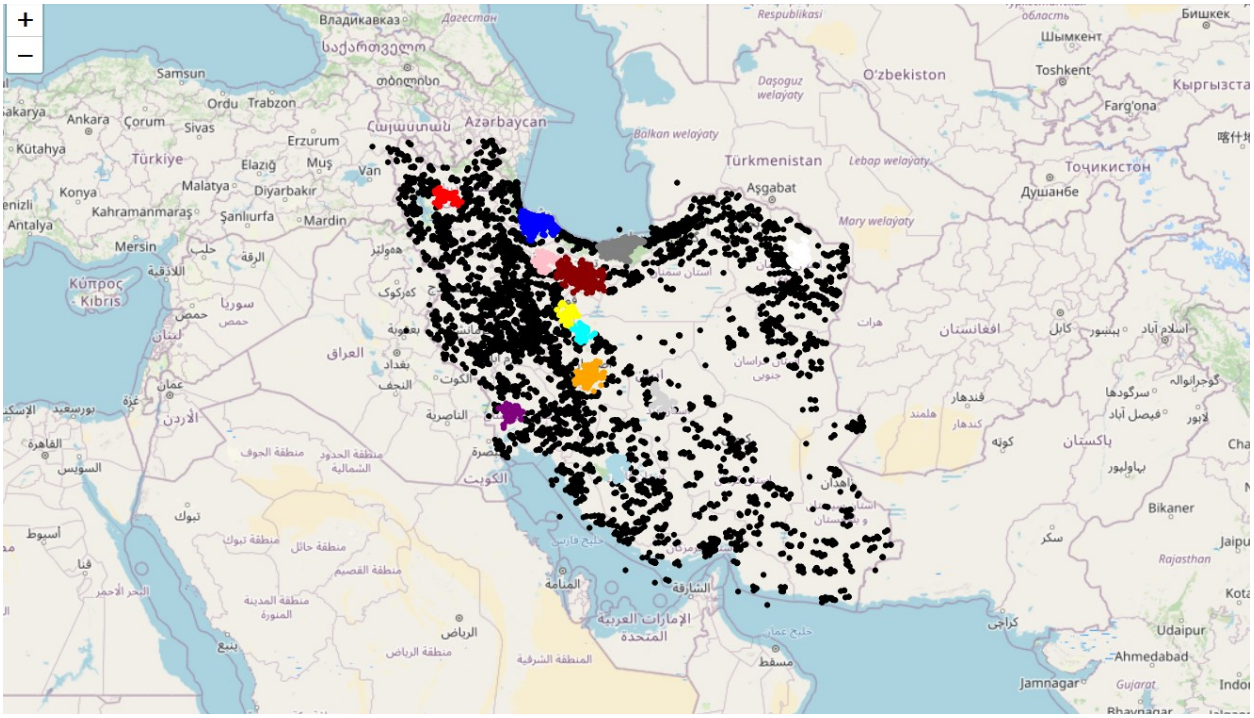
قسمت A :

قسمت B :

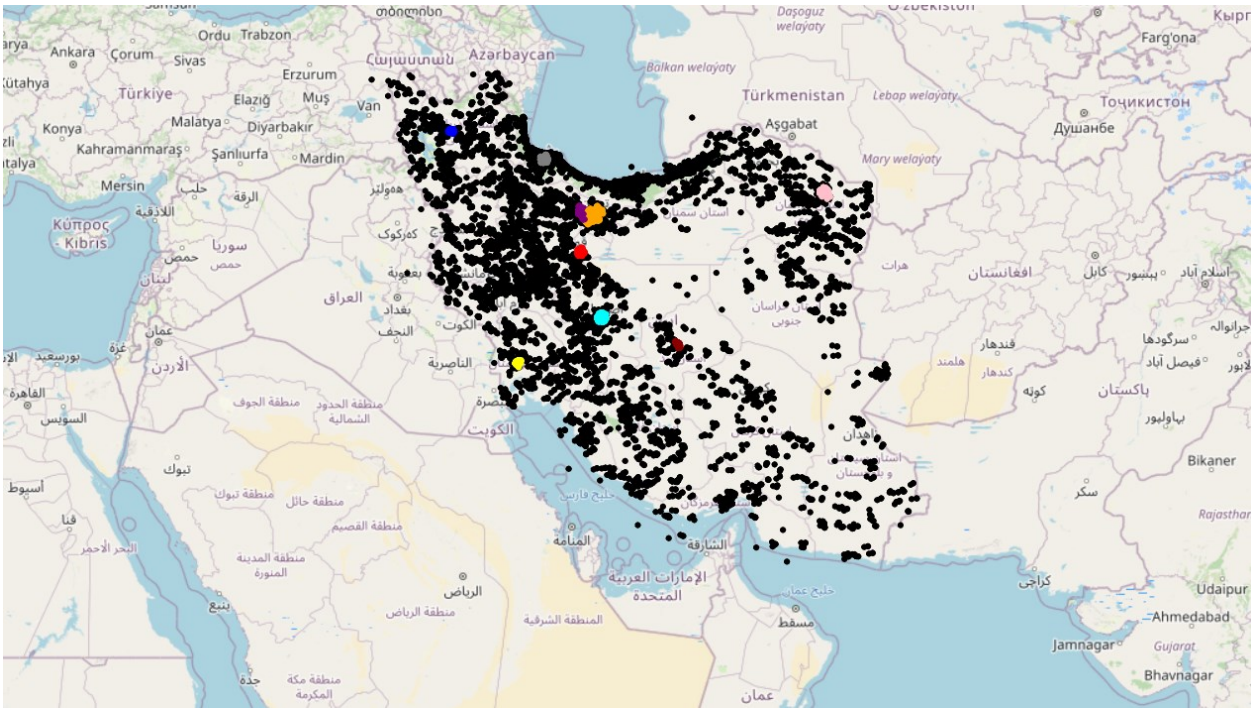
شکل به دست‌آمده با minpts = 200 و eps = 0.5



شکل به دست‌آمده با minpts = 200 و eps = 0.3

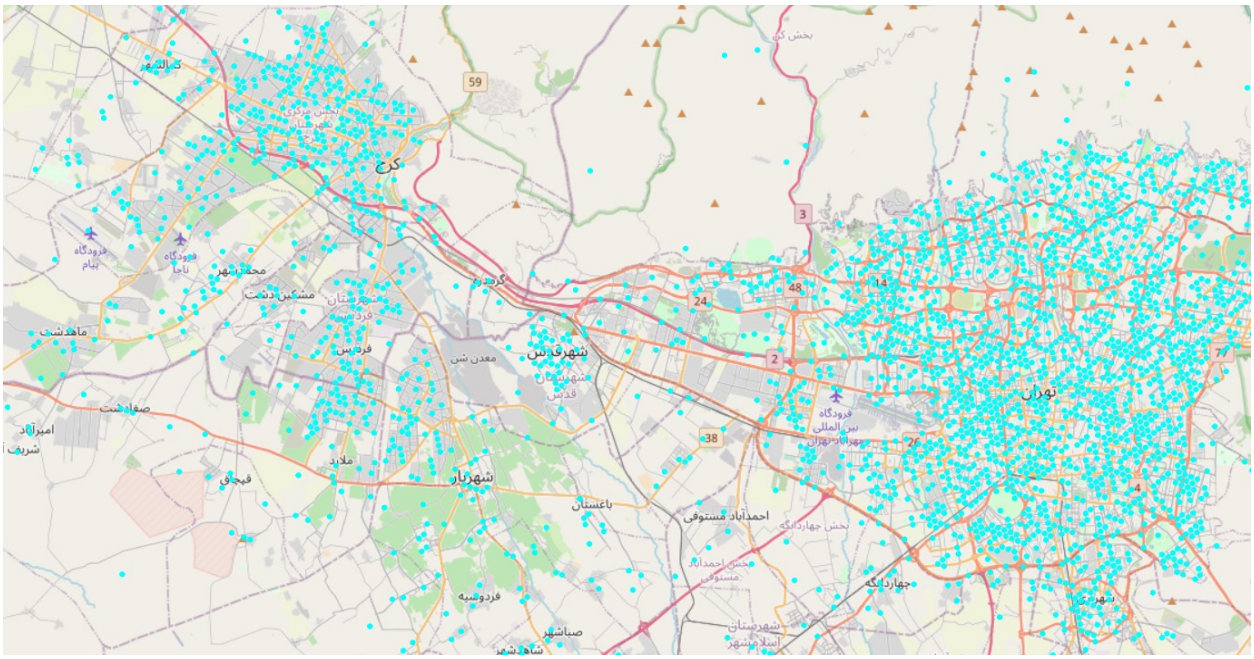


شکل به دست‌آمده با minpts = 200 و eps = 0.1



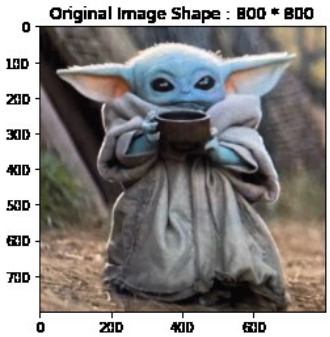
قسمت C و D :

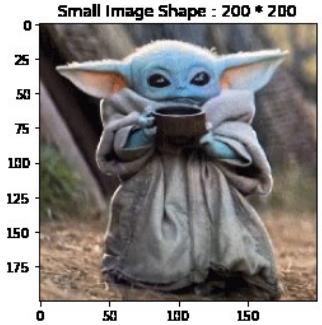
نتایج برای شهرهای کرج و تهران به شکل زیر هستند:



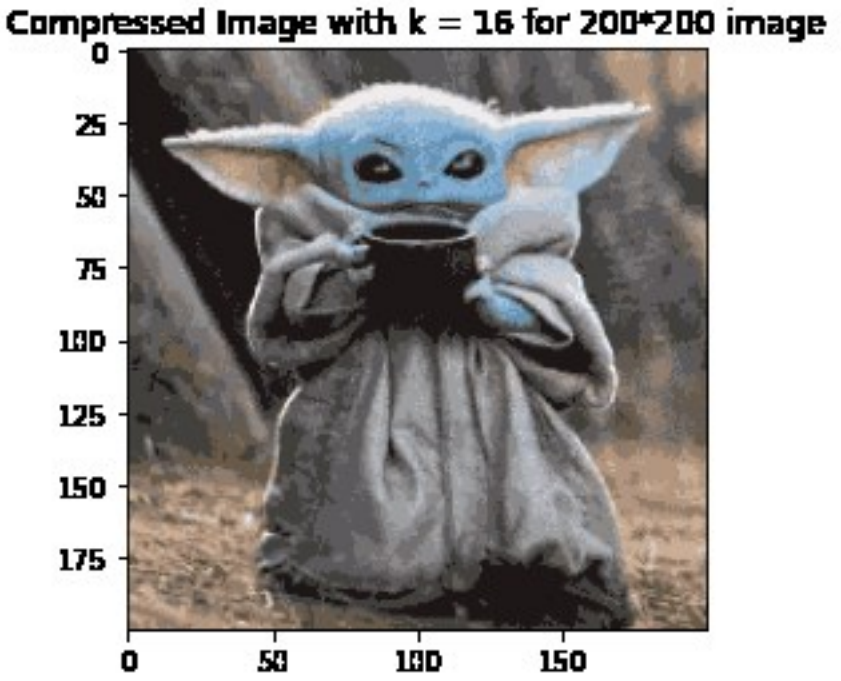
پیاده سازی سوم ؛ فشرده‌سازی عکس

در این بخش می‌خواهیم تصویری با الگوریتمی به نام LUT رویکرد این است که همه رنگ‌های موجود در تصویر را به صورت دسته‌هایی طبقه‌بندی کرده و مرکز ثقل هر خوشه را انتخاب کنند. برای پیاده‌سازی این روش از روش k-means که در بخش اول به عنوان روش خوشه‌بندی استفاده شده‌است، بهره می‌بریم. همچنین می‌دانیم که تصویر یک آرایه سه‌بعدی است که در آن دو بعد اول ارتفاع و عرض تصویر و بعد سوم , مقدار RGB هر پیکسل تصویر است. در ادامه عملکرد این روش را برای تصاویر ورودی کوچک و اولیه آزمون خواهیم کرد و نتایج فشرده‌سازی را از طریق $k = 16$ و $k = 256$ برای هر تصویر نشان می‌دهیم. قبل از خوشه‌بندی ما بررسی می‌کنیم که چگونه بسیاری از رنگ‌ها با استفاده از روش تفاضل محدود در آرایه‌های تصاویر وجود دارند.

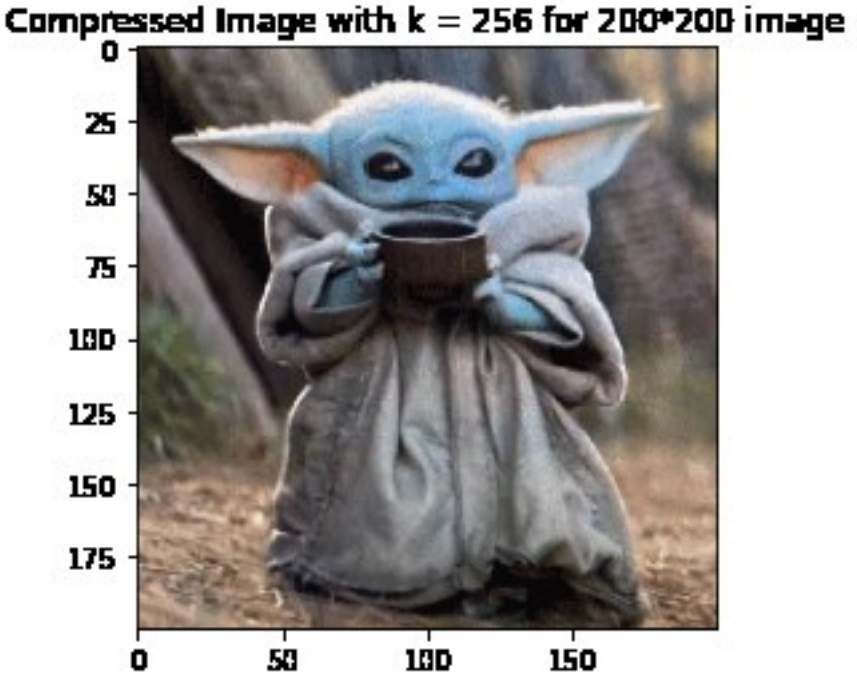




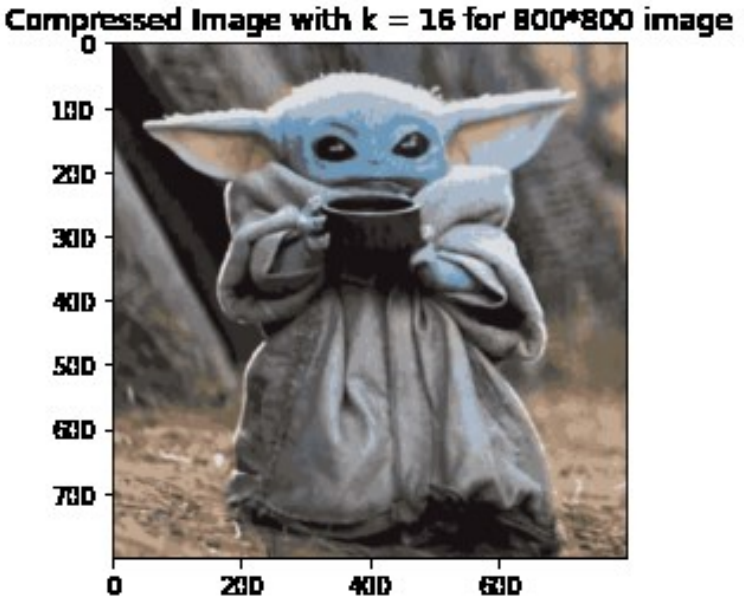
شکل‌های بدست آمده با توجه به خواسته های مسئله به شکل زیر است:



عکس کوچکتر با k = 16



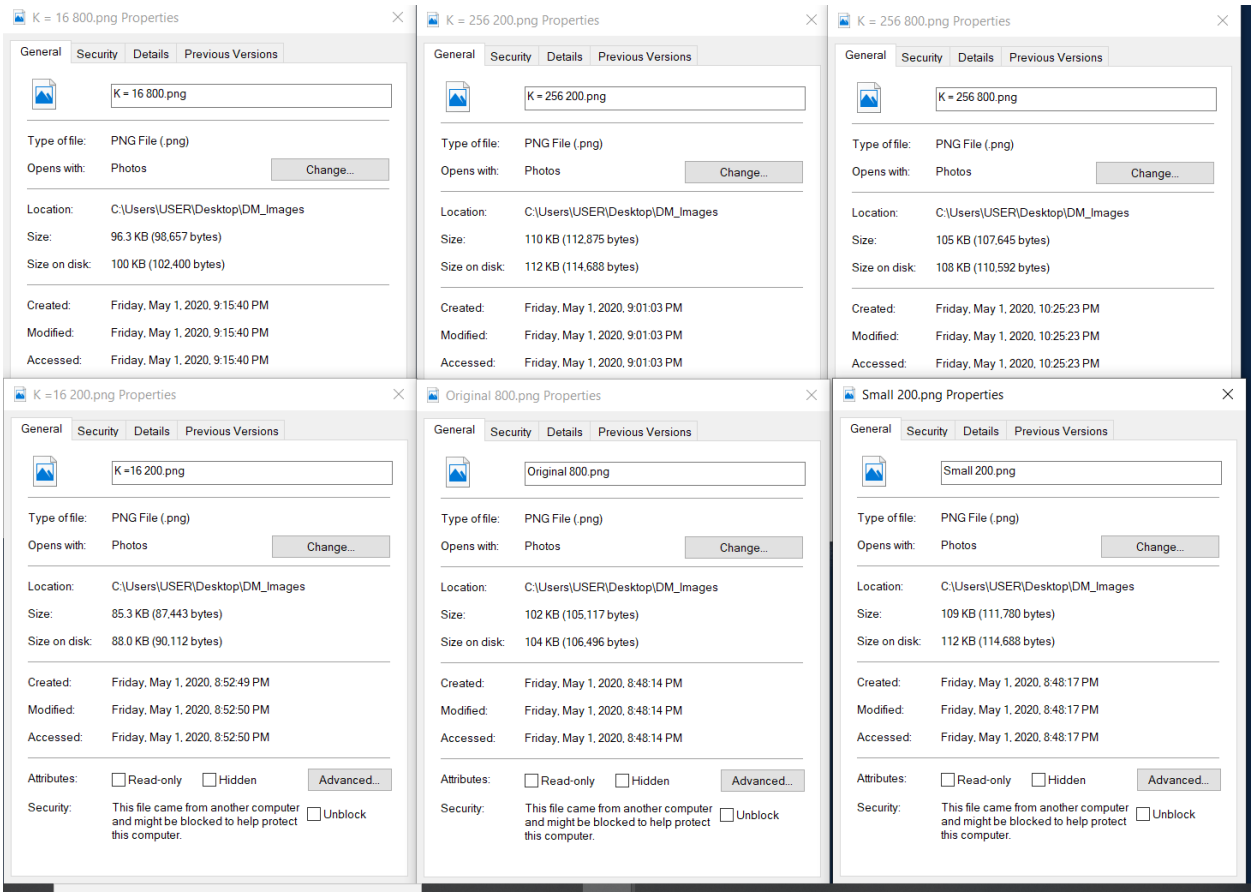
عکس کوچکتر با k = 256



عکس عادی با k = 16

عکس عادی با k = 256

همانطور که می‌توانید ببینید، تعداد دفعات تکرار برای خوشه‌بندی تمام نقاط داده هنگامی که ما از ۲۵۶ سنترویید برای خوشه کردن رنگ‌ها استفاده می‌کنیم، به این دلیل که مقادیر زیادی از هم پوشانی در رنگ‌های مشابه وجود دارد(چون فقط ۱۹۵ رنگ در تصویر وجود دارد.)



در تصویر بالا شما می‌توانید اطلاعات تصاویر ذخیره‌شده در مقایسه با تصویر اصلی را ببینید.

پایان