



دانشگاه خوارزمی

دانشکده فنی و مهندسی

گروه مهندسی صنایع

بخش بندی مشتریان با استفاده از الگوریتم های k-میانگین و SOM

تهیه کننده:

مهدی کشاورز

عنوان درس:

تحلیل داده های مهندسی

در این پژوهش تلاش شده تا با اجرای تکنیک های داده کاوی بر روی دیتابیسی متشکل از داده های مربوط به تراکنش های فروش و داده های مربوط به سفارشات برگشت خورده یک شرکت، مشتریان آن را به بخش های مختلفی تقسیم کرده تا برنامه ریزی و تصمیم گیری درباره هر کدام منظم تر و هدفمند تر شود. پژوهش در ۴ گام معرفی و آماده سازی داده ها، خوشبندی k-میانگین، خوشبندی با SOM و مقایسه نتایج تهیه شده است.

دیتابیس متشکل از حدود ۵۰ هزار رکورد و شامل ویژگی هایی از جمله تعداد خرید، تخفیف، سود، نوع محصول خریداری شده، منطقه جغرافیایی مشتری و نوع ارسال است که پس توضیح و مصوّر سازی داده ها، عملیات آماده سازی بر روی آن ها صورت گرفته است. برای آماده سازی داده ها تلاش شده داده های مفقود به صورت مناسب با میانگین مربوط به ویژگی خودشان جایگزین شده و پس از یکپارچه سازی داده ها، عملیات کاهش ابعاد صورت گیرد تا هم نتایج مقبول تر شود و هم محاسبات آینده سریع تر صورت بگیرد و نتیجه، انباره داده ای متشکل از داده های مناسب داده کاوی است. برای بخش بندی مشتریان دو رویکرد وجود دارد که در اولین مورد با استفاده از خوشبندی k-میانگین مشتریان به چند خوشبندی تقسیم خواهند شد که هر کدام از این خوشبندی ها معرف یک بخش از مشتریان است. پیش از اجرای عملیات خوشبندی از تمایل آماری داده ها به خوشبندی اطمینان حاصل شده و سپس تعداد بهینه خوشبندی ها با استفاده از روش Elbow بدست آمده است. رویکرد دوم مشابه رویکرد اول است با این تفاوت که پیش از خوشبندی، داده ها با استفاده از نقشه های خود سازمانده به داده های ۲ بعدی تبدیل شده اند. این کاهش ابعاد به افزایش دقت خوشبندی ها و رسیدن به نتایج بهتر کمک شایانی می کند. در انتهای پژوهش، نتایج حاصل از دو رویکرد با هم مقایسه شده اند.

فهرست مطالب

۱	مقدمه
۳	۱- گام اول: جمع آوری و آماده سازی داده ها
۴	۱-۱- معرفی پایگاه داده
۴	۱-۲- پیش پردازش داده ها
۵	۱-۲-۱- درک ماهیت صنعت
۵	۱-۲-۲- درک ماهیت داده ها
۸	۱-۲-۳- آماده سازی داده ها
۱۵	۱-۳-۱- مدلسازی
۱۶	۱-۳-۲- ارزیابی
۱۶	۱-۳-۳- اجرا
۱۶	۱-۳-۴- تمایل به خوش بندی
۱۸	۱-۴- تعداد بهینه k
۱۹	۲- گام دوم: خوش بندی با استفاده از K -میانگین
۲۰	۲-۱- ساخت و اجرای مدل
۲۳	۲-۲- تحلیل نتایج به دست آمده
۳۱	۲-۳- مدل های دیگر
۳۸	گام ۳: نقشه خودسازمانده (SOM)
۳۹	۳-۱- مدلسازی
۴۱	۳-۲- نتایج مدل
۵۱	۳-۳- مدل های دیگر
۵۱	۳-۱-۳-۱- مدل به همراه متغیرهای کیفی
۵۵	۳-۲-۳-۲- مدل با سه خوش
۶۲	۴- مقایسه نتایج دو مدل
۶۷	پیوست ها

امروزه انبوه داده^۱ و رشد نمایی آن ها از نظر تعداد و حجم، بلای جان شرکت ها و کسب و کارهای مختلف شده است و ذخیره سازی آن ها علاوه بر روش های اصولی، نیازمند فضای سخت افزاری زیادی است. داده کاوی^۲، کوششی است برای کسب دانش^۳ از این داده های خام و از جمله علومی است که اهمیت آن روز به روز بیشتر و آشکارتر می شود. داده کاوی دانشی است برای استخراج قواعد، روابط و الگوهای مشخص در میان داده های خام و استفاده نشده و فرآیندی است که به وسیله آن، از داده به اطلاعات^۴ می رسیم. در نهایت، این اطلاعات وسیله ای برای کسب دانش و به دنبال آن خرد سازمانی است.

امروزه داده کاوی در علوم مختلف جایگاه ویژه ای یافته و بسیاری از پژوهشگران، از تکنیک های آن برای اهداف مختلف استفاده می کنند. مهندسان فرآیند با اعمال تکنیک ها داده کاوی به دنبال پیش بینی خرابی های آینده در صنایع مختلف هستند؛ محققان حوزه وب، از داده کاوی برای استخراج اطلاعات مناسب از کاربران وب سایت ها و افزایش نرخ تبدیل کسب و کارشان استفاده می کنند؛ پژوهشگران حوزه سلامت، از داده های بیماران برای پیش بینی بیماری های آینده آنان و پیش بینی بیماری های دیگر افراد استفاده می کنند. داده کاوی به دلیل اهمیتی که در آشکارسازی بسیاری از اطلاعات دارد، در رشته های مختلف کاربرد روزافزونی دارد.

فرآیند داده کاوی و کشف دانش^۵ با استفاده از تکنیک های مختلفی صورت می گیرد. پیش از به کارگرفتن هر کدام از این تکنیک ها، لازم است داده ها آماده شده و سپس بسته به ماهیت داده ها و هدف از داده کاوی، تکنیک مناسب اعمال شود. از جمله تکنیک هایی که علاوه بر سادگی مزیت های دیگری نیز به همراه دارد، تکنیک خوشه بندی^۶ است که داده ها بر مبنای ویژگی های^۷ آنان و نزدیکی این ویژگی ها به هم، در یک خوشه^۸ قرار می گیرند. خوشه بندی در واقع کوششی است برای قرار دادن داده های مشابه در کنار هم بنابراین اگر میان داده ها تفاوتی نباشد، خوشه بندی بی فایده است.

خوشه بندی می تواند برای اهداف گوناگونی به کار رود که یکی از مهم ترین آن ها، بخش بندی مشتری ها برای شرکت ها می باشد. این تقسیم بندی به شرکت ها در تصمیم گیری و برنامه ریزی کمک زیادی می کند زیرا می توانند مشتری های خود را بر مبنای ویژگی های مشخصی در یک بخش قرار داده و در نتیجه برای هر بخش، رویکرد مناسبی به کار بگیرند. برای مثال ممکن است برخی از مشتریان به دنبال کیفیت باشند و حاضر باشند بهای آن را نیز بپردازند و یا برخی دیگر تنها به دنبال صرفه اقتصادی باشند و یا گروهی دیگر در جست و جوی تعادلی میان کیفیت و هزینه باشند. بدیهی است که نمی توان همه آنان را با یک چشم دید و لازم است برای هر کدام رویکردی مناسب به کار گرفت. حتی شرکت هایی نیز که بر بخش بسیار خاصی از بازار تمرکز دارند باز هم با اختلاف در سلیقه و دیگر ویژگی های مشتریان روبرو می باشند.

¹ Data

² Exponential

³ Data Mining

⁴ Knowledge

⁵ Information

⁶ Knowledge Discovery

⁷ Clustering

⁸ Features

⁹ Cluster

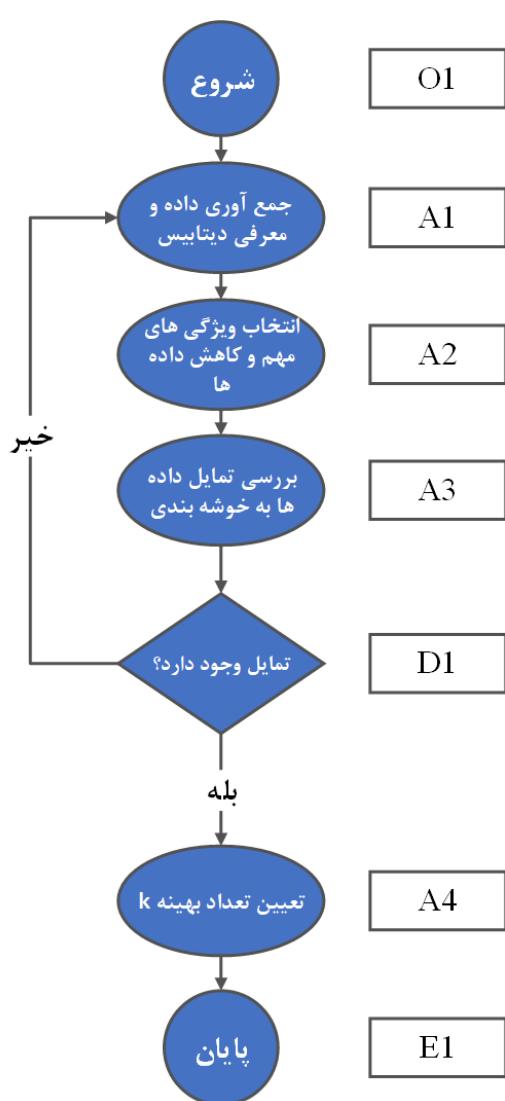
در ادامه، قصد داریم مشتریان یک شرکت را با استفاده از پایگاه داده تراکنش های آنان خوشه بندی کنیم. فرآیند در ۴ گام صورت می‌گیرد که در گام ۱، به معرفی پایگاه داده و آماده سازی داده ها می‌پردازیم. همچنین ویژگی های مهمی که لازم است مورد بررسی قرار بگیرند و انتخاب و از بقیه صرف نظر می‌کنیم. در این گام همچنین تمایل داده ها به خوشه بندی را بررسی کرده و تعداد بهینه خوشه ها را محاسبه می‌کنم. در گام ۲، فرآیند خوشه بندی با استفاده از روش k -میانگین صورت می‌گیرد. در گام ۳، داده کاوی با تکنیک متفاوتی به نام نقشه های خودسازمانده^۱ یا SOM انجام می‌شود(همچنین به کمک خوشه بندی k -میانگین) که ابزاری برای کاهش ابعاد داده به ۲ بعد و مصور سازی خوشه های حاصل است. در گام ۴ و پایانی، نتایج حاصل از ۲ روش با هم مقایسه شده و مزایا و معایب هر کدام ذکر می‌شود. لازم به ذکر است که در هر گام، فرآیند با فلوچارت نشان داده شده و توضیحات جزئی تر درباره هر تکنیک و مراحل آن داده می‌شود. همچنین نتایج حاصل از هر روش در گام مربوط به هر کدام به صورت جداگانه بررسی شده است. در بخش پیوست ها نیز، توضیحات فایل هایی که ضمیمه این پژوهش می‌باشند بیان شده است.

¹ Self-Organizing Map

۱- گام اول: جمع آوری و آماده سازی داده ها

پیش از شروع، لازم به ذکر است که در این پژوهش از نسخه ۹,۷ نرم افزار RapidMiner برای داده کاوی استفاده شده است. در یک مورد نیز از زبان پایتون کمک خواهیم گرفت. پایگاه داده مورد استفاده، یک فایل اکسل با فرمت **xlsx** می باشد و با تغییر فرمت به شکل **CSV** درآمده و نرم افزار RapidMiner به خوبی از آن پشتیبانی می کند.

برای داشتن رویکردی منطقی و منظم، از یک فلوچارت برای نمایش مراحل هر گام از پژوهش استفاده خواهیم کرد. فلوچارت این گام در شکل ۱ آورده شده است. مراحل این فلوچارت و فلوچارت های دیگر کدگذاری شده تا بتوان در صورت نیاز به آن ها ارجاع داده و توضیحات تکمیلی را بیان کرد.



A3: در این مرحله، از هیستوگرام مسافت های زوجی^۱ استفاده می شود.

D1: در صورتی که نمودار بدست آمده دارای ۲ قله باشد، نشان از تمايل داده ها به خوش بندی است.

A4: برای انتخاب تعداد بهینه خوش ها، از روش ELBOW استفاده می شود. در این روش، آزمایش را با تعداد مختلف K انجام می دهیم و میانگین را در فاصله مراکز ثقل برای هر K نمودار می کنیم. در نقطه ای از نمودار که شکست پیدید آمد، تعداد مناسب خوش به دست آمده است.

شکل ۱ – فلوچارت خوش بندی با روش K-میانگین

¹ Pairwise Distance

۱-۱- معرفی پایگاه داده

در ابتداء، پایگاه داده با توجه به موضوع مورد نظر جمع آوری شده و عملیات آماده سازی بر روی آن ها صورت می‌گیرد. پایگاه داده مورد استفاده، تراکنش های فروش یک شرکت را با بیش از ۵۰ هزار رکورد در یک فایل اکسل نشان می‌دهد که در آن اطلاعات مشتریان، نوع خرید، نوع ارسال، منطقه جغرافیایی و ... آورده شده است. در این فایل ۳ صفحه (sheet) وجود دارد که از ۲ صفحه اول آن استفاده خواهد شد. صفحه اول داده های مربوط به سفارشات با ۵۱۲۹۰ رکورد و ۲۴ ویژگی است. ویژگی ها به ترتیب حضور در پایگاه داده عبارتند از:

مشخصه یا ID سطر، مشخصه یا ID سفارش، تاریخ سفارش، تاریخ ارسال، نوع ارسال، مشخصه یا ID مشتری، نام مشتری، بخش^۱، کد پستی، شهر، ایالت، کشور، منطقه (برای مثال اقیانوسیه، آفریقای غربی و...)، بازار، مشخصه یا ID محصول، دسته بندی^۲ یا مجموعه محصول، زیرمجموعه محصول، نام محصول، فروش، تعداد، تخفیف، سود، هزینه ارسال و اولویت ارسال.

صفحه دوم، اطلاعات مربوط به سفارشات برگشت خورده می‌باشد که دارای ۱۰۸۰ رکورد و ۳ ویژگی زیر است:

برگشت، شناسه سفارش، منطقه

خوشه بندی مشتریان بر مبنای سفارش ها (داده های موجود در تراکنش ها) صورت خواهد گرفت. خوشه بندی و در نهایت بخش بندی^۳ مشتریان بر مبنای روش های سنتی و اطلاعاتی موجود از پیش، دانش خاصی را آشکار نخواهد ساخت. برای مثال بخش بندی به بازار اروپا، اقیانوسیه و آمریکا یکی از این راه ها است اما اطلاعات مناسبی آشکار نخواهد کرد و در پاسخ به پرسش های مانند پرسش های زیر ناتوان خواهیم بود:

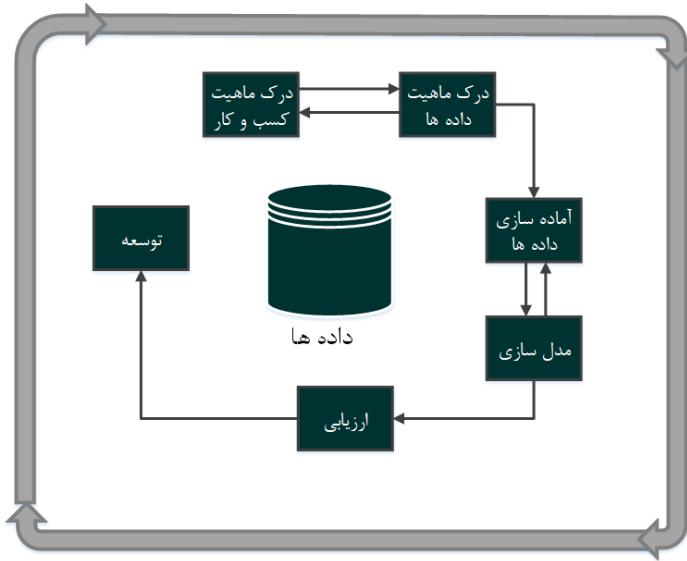
چند درصد از تراکنش های مربوط به اروپا سود آور بوده؟ الگوی خرید در آمریکا چیست؟ بر چه مبنایی می‌توان مشتری های هر منطقه جغرافیایی را در یک خوشه قرار داد؟

برای پاسخ به این پرسش ها، لازم است عملیات داده کاوی با استفاده از تکنیک خوشه بندی صورت بگیرد.

۱-۲- پیش پردازش داده ها

پس از معرفی پایگاه داده، عملیات آماده سازی داده ها صورت می‌گیرد که وقت گیرترین بخش از فرآیند داده کاوی می‌باشد. متدولوژی CRISP برای آماده سازی داده ها استفاده خواهد شد. متدولوژی CRISP در شکل صفحه بعد قابل مشاهد است.

1 Segment	2
1 Category	3
^۱ Segmentation	4



شکل ۲ – متدولوژی CRISP

رویکرد مذکور در ادامه به صورت گام به گام اجرا و توضیح داده شده است.

۱-۲-۱- درک ماهیت صنعت

داده ها، مربوط به فروش یک شرکت است که محصولات آن در ۳ دسته بندی قرار می‌گیرند. لوازم خانگی، تجهیزات اداری و تجهیزات تکنولوژیک(دستگاه کپی، تلفن و...). فروش به صورت جهانی است و مشتری های زیادی وجود دارد که برای هر کدام داده های وسیعی ثبت شده است. بنابراین لازم است که آماده سازی به نحوی باشد که خوشه بندی را برای حجم زیادی از داده ها هموار کند.

۱-۲-۲- درک ماهیت داده ها

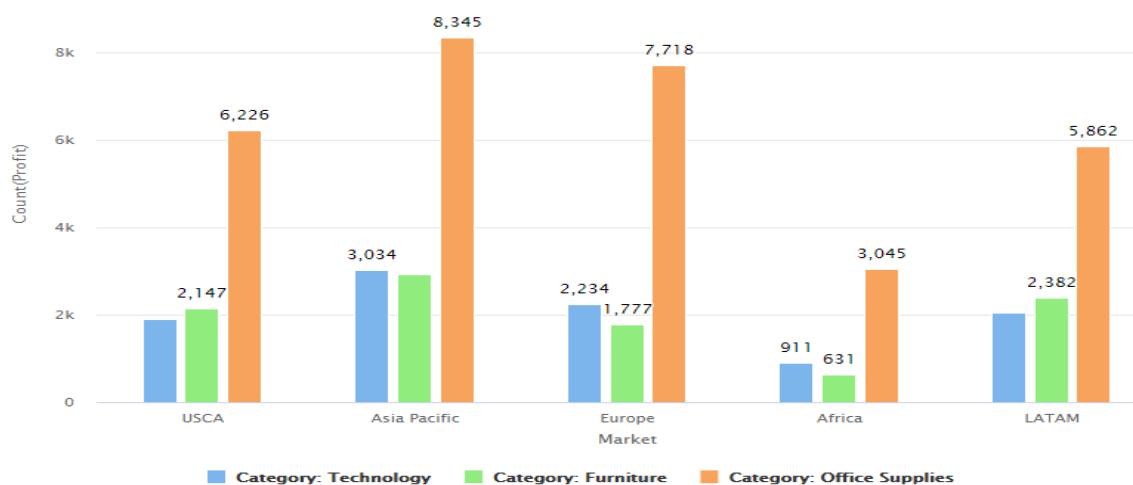
از آن جا که داده ها تراکنش های فروش را نشان می‌دهند، می‌توانند گزینه مناسبی برای خوشه بندی مشتریان بر مبنای رفتار خرید آنان باشند. تعداد ابعاد داده ها به گونه ای است که ویژگی های مهمی از داده ها را در بر می‌گیرد؛ برای مثال تعداد خرید، محصول خریداری شده، نوع بازار و ... بنابراین کفیت داده ها برای بررسی رفتار مشتریان مناسب می‌باشد.

پیش از آماده سازی داده ها، با استفاده از تلخیص توصیفی داده ها و استفاده از نمودار، اطلاعاتی پیش زمینه ای درباره داده ها بدست خواهیم آورد. این رویکرد پیش از شروع عملیات آماده سازی داده ها، به درک عمیق تر داده ها و مدلسازی مناسب کمک زیادی می‌کند.

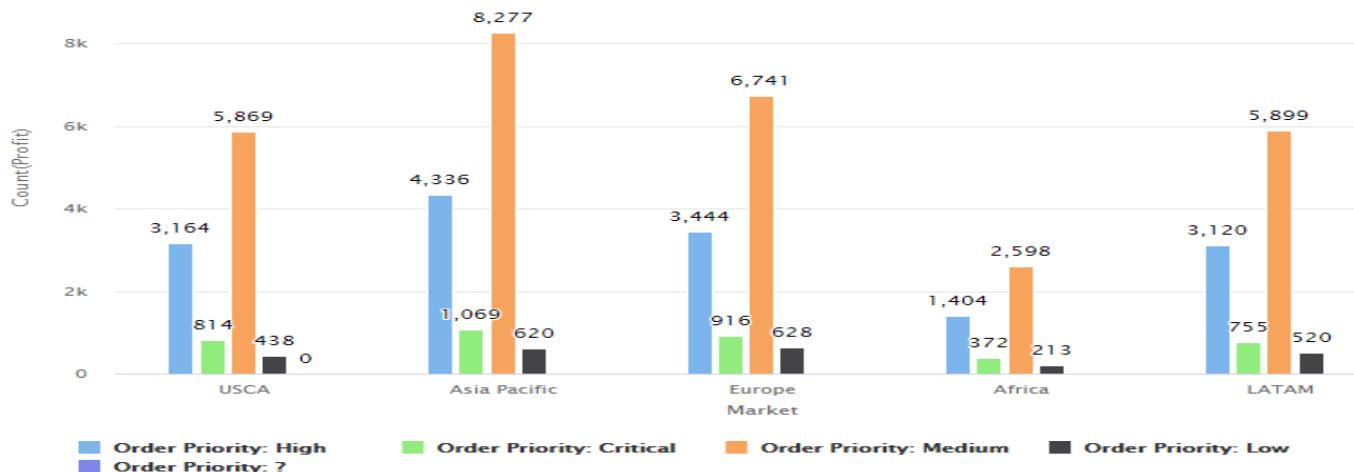
جدول ۱ - خلاصه اطلاعات کمی دیتابیس

میانگین	حداکثر	حداقل	مشخصه
۲۴۶,۸۷۹	۲۲۶۳۸,۴۸۰	۰,۴۴	فروش
۳,۴۷۶	۱۴	۱	تعداد
۰,۱۴۳	۰,۸۵	۰	تحفیف
۲۸,۶۴۱	۸۳۹۹,۹۸۰	- ۶۵۹۹,۹۸۰	سود
۲۶,۵۲۰	۹۳۳,۵۷۰	۱	هزینه ارسال

با استفاده از شکل های زیر، داده ها به صورت اشکال و نمودارهای قابل تفسیر نشان داده شده است.

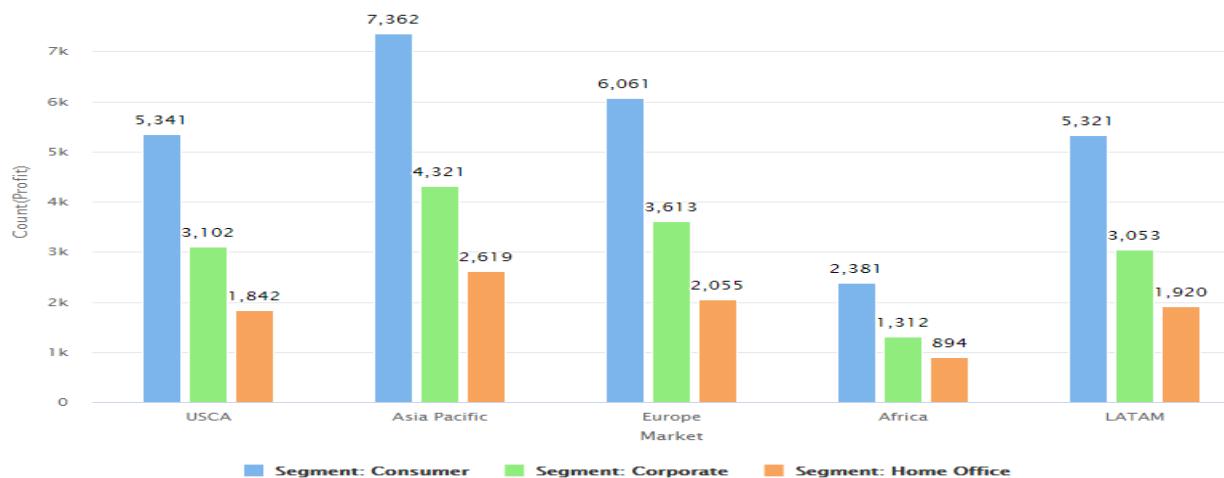


شکل ۳ - میزان سودآوری بازارها و دسته بندی محصولات

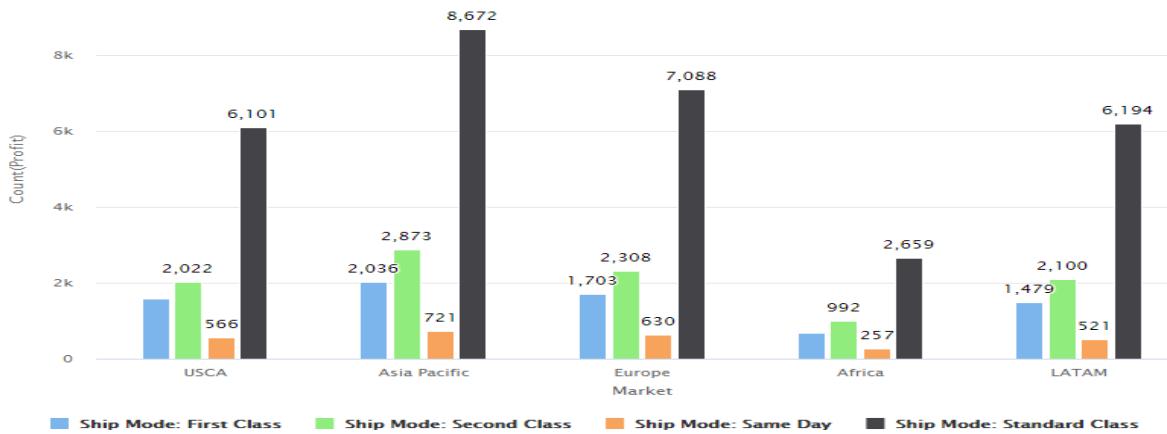


شکل ۴ - میزان سودآوری بازارها و اولویت های ارسال

از شکل ۴ می توان متوجه شد که در دیتابیس داده های مفقود وجود دارد.

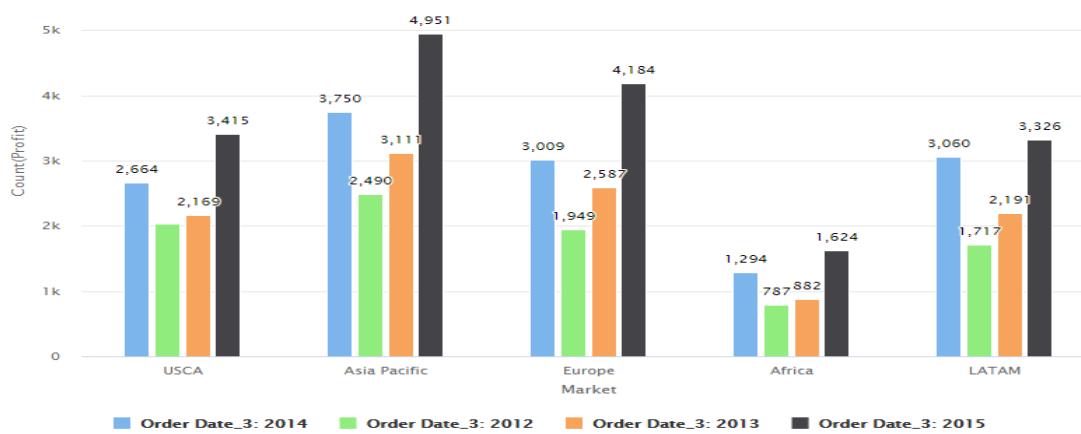


شکل ۵ – میزان سودآوری بازارها و بخش ها



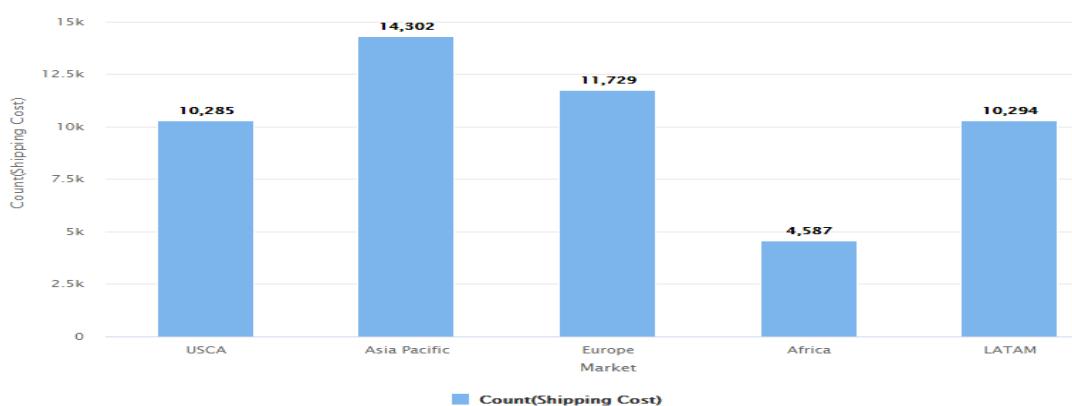
شکل ۶ – میزان سودآوری بازارها و نوع ارسال

از نمودارهای بالا می توان متوجه شد که دسته بندی تجهیزات اداری، اولویت ارسال متوسط، بخش مشتری و نوع ارسال استاندارد بیشترین بخش سود را تشکل می دهد.



شکل ۷ – میزان سودآوری بازارها و تاریخ سفارش

همچنین به نظر میرسد سال ۲۰۱۵ به نسبت دیگر سال ها سودآوری بیشتری به همراه داشته است.



شکل ۸ – هزینه ارسال بر اساس بازارهای مختلف

در نهایت، از شکل ۸ می‌توان مشاهده کرد که هزینه‌های ارسال چگونه میان بازارهای مختلف توزیع شده است. بازار آسیا-اقیانوسیه در این نمودار بیشترین هزینه را دارد. توجه شود که همه این اعداد در این نمودار و نمودارهای پیشین به صورت مجموع نوشته شده‌اند.

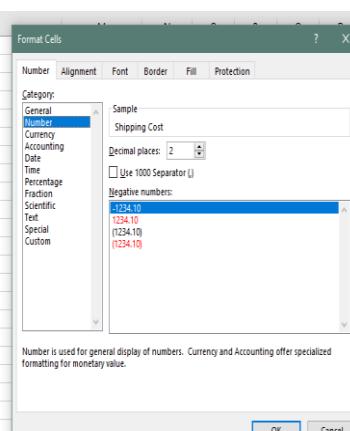
۱-۲-۳- آماده سازی داده‌ها

برای آماده سازی داده‌ها، ۵ گام زیر صورت می‌گیرد.

۱-۲-۱- پاکسازی داده‌ها

برای پاکسازی داده‌ها، لازم است درباره داده‌های مفقود، مغلوش، متناقض، عمدى و ناسازگار تصمیم گیری شود. در پایگاه داده، داده‌های مفقود بسیاری به چشم می‌خورد. بیشتر این مقادیر مفقود، مربوط به ویژگی کدپستی می‌باشند که در فایل اکسل با نام **blank** شناخته می‌شوند(در سلول **Combo**). در ادامه توضیح داده خواهد شد که ویژگی کد پستی به دلیل بی اهمیت بودن حذف خواهد شد و تاثیری در خوشبندی نخواهد گذاشت.

در ستون مربوط به سود و ویژگی‌های کیفی مانند کشور و منطقه نیز داده مفقود وجود دارد. پیش از آنکه رویکرد برخورد با این داده‌ها را بیان کنیم، ابتدا در فایل اکسل، ستون مربوط به هزینه‌ها و درآمد‌ها را انتخاب کرده و نوع داده‌های آن را به داده عددی تبدیل می‌کنیم تا علاوه بر پاک شدن علامت دلار در مقابل آن‌ها، نرم افزار نیز در شناسایی نوع این ویژگی‌ها خطأ نداشته باشد.



R	S	T	U	V	W	X	Y
Product	Sales	Quantity	Discount	Profit	Shipping	Order Priority	
Samsung C	\$221.98	2	0	\$62.15	40.77	High	
Novimex I	\$3,709.40	9	0.1	-\$288.77	923.63	Critical	
Nokia Smz	\$5,175.17	9	0.1	\$919.97	915.49	Medium	
Motorola S	\$2,892.51	5	0.1	-\$96.54	910.16	Medium	
Sharp Wtr	\$2,832.96	8	0	\$311.52	903.04	Critical	
Samsung C	\$2,862.68	5	0.1	\$763.28	897.35	Critical	
Novimex I	\$1,822.08	4	0	\$564.84	894.77	Critical	
Chromcraft	\$5,244.84	6	0	\$996.48	878.38	High	
Sauder Fa	\$341.96	2	0	\$54.71	25.27	High	
Global Pu	\$48.71	1	0.2	\$5.48	11.13	High	
Newell 33	\$17.94	3	0	\$4.66	4.29	High	
Bevis Con	\$4,626.15	5	0	\$647.55	835.57	High	
Cisco Sma	\$2,616.96	4	0	\$1,151.40	832.41	Critical	
Harbour C	\$2,221.80	7	0	\$622.02	810.25	Critical	
KitchenAi	\$3,701.52	12	0	\$1,036.08	804.54	Critical	
Breville R	\$1,869.59	4	0.1	\$186.95	801.66	Critical	
Akro Stack	\$12.62	2	0.2	-\$2.52	1.97	Low	
Hoover St	\$7,958.58	14	0	\$3,979.08	778.32	Low	
Brother Fc	\$2,565.59	9	0.1	\$28.40	766.93	Critical	
KitchenAi	\$3,409.74	6	0	\$818.28	763.38	High	
Hon Comp	\$1,977.72	4	0	\$276.84	759.47	Critical	
Carina 42"	\$242.94	3	0	\$4.86	1.28	High	

S	T	U	V	W	X	
Sales	Quantity	Discount	Profit	Shipping	Order Priority	
221.98	2	0	62.15	40.77	High	
3709.40	9	0.1	-288.77	923.63	Critical	
5175.17	9	0.1	919.97	915.49	Medium	
2892.51	5	0.1	-96.54	910.16	Medium	
2832.96	8	0	311.52	903.04	Critical	
2862.68	5	0.1	763.28	897.35	Critical	
1822.08	4	0	564.84	894.77	Critical	
5244.84	6	0	996.48	878.38	High	
341.96	2	0	54.71	25.27	High	
48.71	1	0.2	5.48	11.13	High	
17.94	3	0	4.66	4.29	High	
4626.15	5	0	647.55	835.57	High	
2616.96	4	0	1151.40	832.41	Critical	
2221.80	7	0	622.02	810.25	Critical	
3701.52	12	0	1036.08	804.54	Critical	
1869.59	4	0.1	186.95	801.66	Critical	
12.62	2	0.2	-2.52	1.97	Low	
7958.58	14	0	3979.08	778.32	Low	
2565.59	9	0.1	28.40	766.93	Critical	
3409.74	6	0	818.28	763.38	High	
1977.72	4	0	276.84	759.47	Critical	
242.94	3	0	4.86	1.28	High	

شکل ۹ - مراحل پاکسازی در اکسل

هنگام وارد کردن داده ها در محیط نرم افزار، نقش هر ویژگی مناسب با ماهیت آن مشخص می شود. برای مثال، ویژگی فروش از نوع **real** و یا ویژگی تعداد از نوع **integer**(عدد صحیح) می باشد.



شکل ۱۰ - عملگر خواندن داده ها

در شکل بالا، داده های موجود در صفحه اول دیتابست وارد شده اند و در مجموعه شکل های زیر، چگونگی تعیین نوع هر کدام از ویژگی ها آورده شده است.

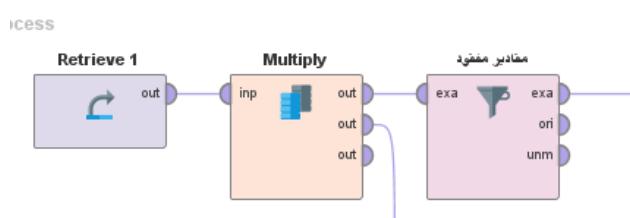
ویژگی های فروش، قیمت و سود از نوع **Real** و ویژگی تعداد از نوع **Integer** انتخاب شده است. ویژگی **Returned** نیز از نوع **Binominal**(دوتایی) و مابقی ویژگی ها، از نوع **Polynomial** می باشند.

توجه: برای اینکه امکان استخراج سال از تاریخ های موجود باشد(حذف روز و ماه و تنها باقی گذاشتن سال)، نوع ویژگی های تاریخی **Polynomial** انتخاب شده است.

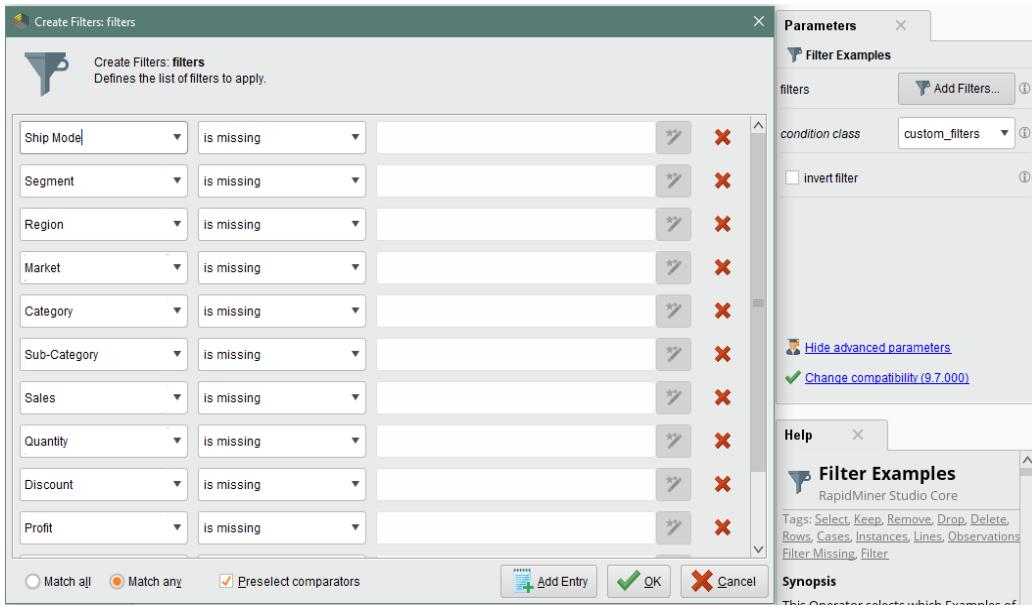
	Customer N... (polynomial)	Segment (polynomial)	Postal Code (polynomial)	City (polynomial)	State (polynomial)	Country (polynomial)
1	Aaron Bergman	Consumer	73120	Oklahoma City	Oklahoma	United States
2	Justin Ritter	Corporate	?	Wollongong	New South Wales	Australia
3	Craig Reiter	Consumer	?	Brisbane	Queensland	Australia
4	Katherine Murray	Home Office	?	Berlin	Berlin	Germany
5	Rick Hansen	Consumer	?	Dakar	Dakar	Senegal
6	Jim Mitchum	Corporate	?	Sydney	New South Wales	Australia
7	Toby Swindell	Consumer	?	Porirua	Wellington	New Zealand
8	Mick Brown	Consumer	?	Hamilton	Waikato	New Zealand
9	Aaron Bergman	Consumer	73120	Oklahoma City	Oklahoma	United States
10	Aaron Bergman	Consumer	98103	Seattle	Washington	United States
11	Aaron Bergman	Consumer	98103	Seattle	Washington	United States

شکل ۱۱ - تعیین نقش ویژگی ها

در ابتدای آماده سازی داده ها، داده های مفقود را با استفاده از عملگرهای مناسب در نرم افزار نشان خواهیم داد. این عمل دقیقا پس از وارد کردن دیتابست انجام می شود تا در همان ابتدا، داده های مفقود شناسایی و بهترین رویکرد در برابر آنان اتخاذ شود.



شکل ۱۲ - استفاده از عملگر فیلتر برای شناسایی داده های مفقود



شکل ۱۳ – جزئیات عملگر فیلتر

در شکل بالا، مهم ترین ویژگی هایی که برای داده کاوی از آن ها استفاده خواهیم کرد نشان داده شده اند. نحوه انتخاب این ویژگی ها در ادامه بحث خواهد شد.

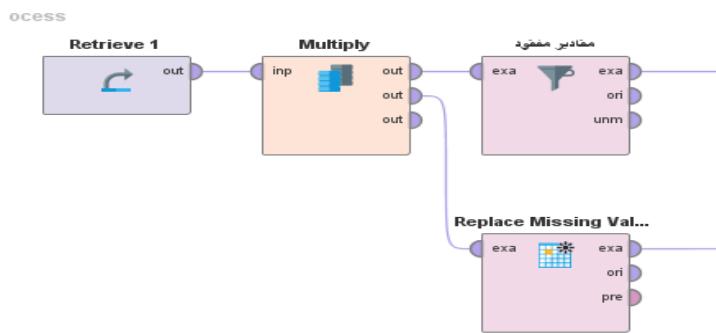
همانطور که در شکل زیر نشان داده شده است، تنها ۱۵ رکورد دارای مقادیر مفقود در ویژگی های مورد نظر بودند که به جز ویژگی اولویت ارسال که کیفی و رتبه ای است، بقیه از نوع کمی می باشند.

Sales	Quantity	Discount	Profit	Shipping Cost	Order Priority
?	?	?	?	?	?
?	?	?	?	?	?
?	?	?	?	?	?
?	?	?	?	?	?
?	?	?	?	?	?
?	?	?	?	?	?
?	?	?	?	?	?
?	?	?	?	?	?
?	?	?	?	?	?
?	?	?	?	?	?
?	?	?	?	?	?
?	?	?	?	?	?
?	?	?	?	?	?
?	?	?	?	?	?
?	?	?	?	?	?

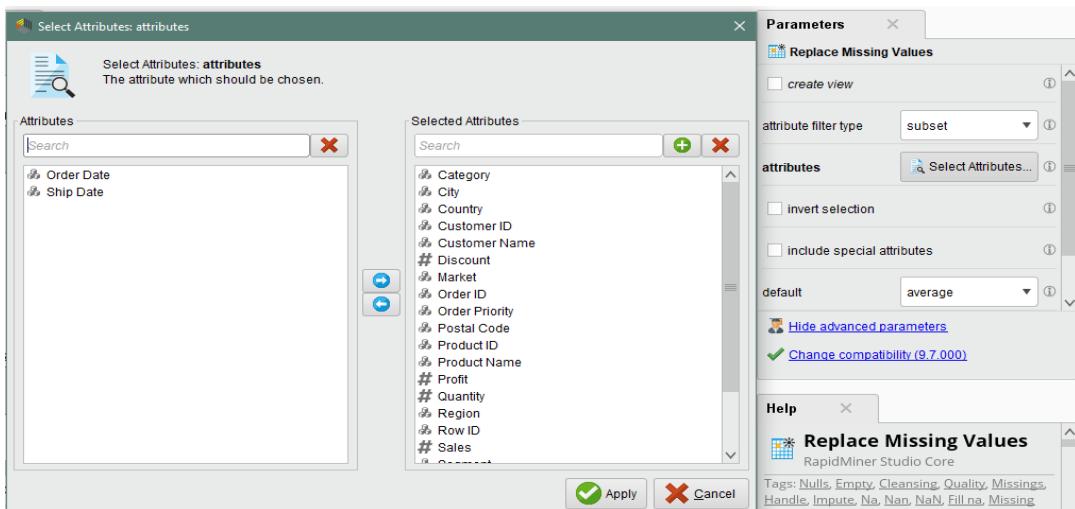
شکل ۱۴ – مقادیر مفقود در ویژگی های مهم

در برابر داده های مفقود، مناسب ترین رویکرد این است که این مقادیر، با میانگین اعداد موجود پر شوند. در گام های ابتدایی توضیح داده شد که هدف ما خوشه بندی مشتری ها بر مبنای رفتار خرید آنان می باشد و در هر صنعت بازه مشخصی از سود وجود دارد بنابراین داده های مفقود نیز از این مورد مستثنی نمی باشند و بهترین رویکرد در قبال آنان، پر شدن با میانگین اعداد دیگر است.

برای این کار، پیش از اجرای مرحله خوشه بندی در نرم افزار و بلافاصله پس از وارد کردن داده ها، مقادیر مفقود را با میانگین دیگر مقادیر پر خواهیم کرد:



شکل ۱۵ – مدل تكميل مقادير مفقود



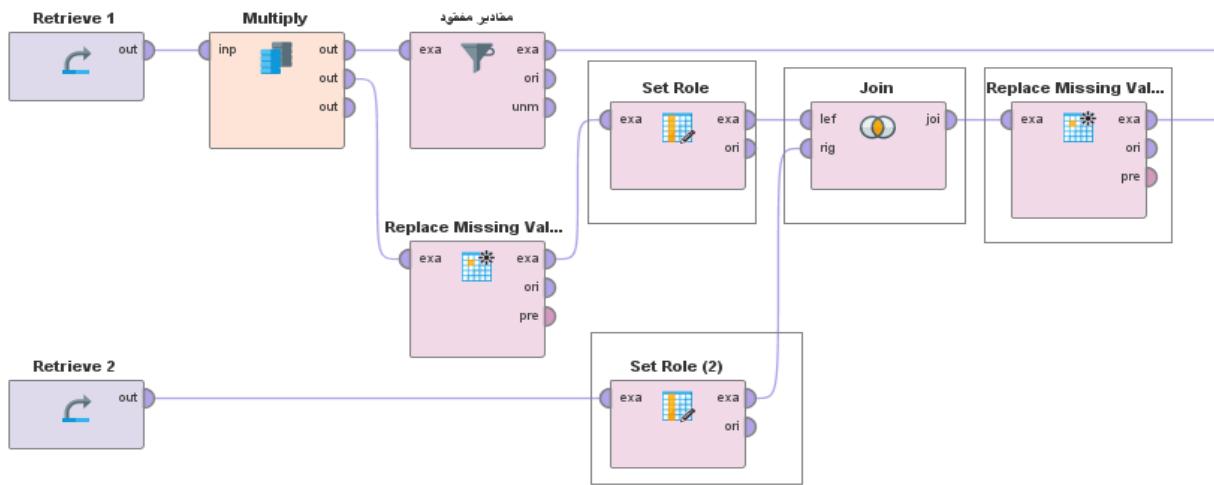
شکل ۱۶ – جزئيات عملگر پر کردن داده های مفقود

در شکل بالا، نشان داده شده که تمامی مقادیر مفقود ویژگی های مورد نظر، با میانگین داده های دیگر پر خواهند شد و قبل تر نیز توضیح داده شد که این رویکرد مناسب است. بسیاری از این مشخصه ها به دلیل بی اهمیت بودن در ادامه حذف خواهند شد. از میان مقادیری که در نهايیت استفاده خواهند شد(ویژگی های مطلوب برای داده کاوی)، جايگزینی تنها بر روی ۱۵ رکورد صورت خواهد گرفت و اين عملیات تاثیر چندانی بر نتایج ندارد اما از نمایش خطأ در نرم افزار جلوگیری می کند.

۱-۲-۳- یکپارچه سازی داده ها

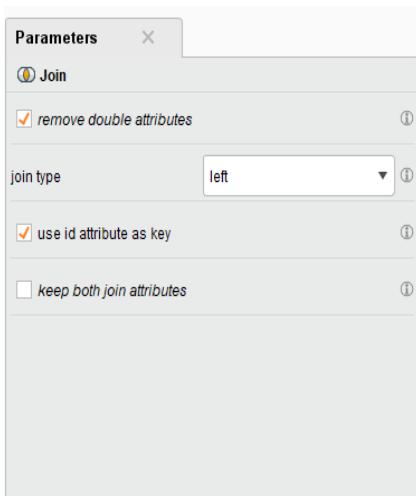
پایگاه داده مورد استفاده شامل سه صفحه يا Sheet می باشد: سفارش ها، مرجوعات و مشتری ها. صفحه مشتری ها کاربرد چندانی ندارد اما لازم است میان صفحه سفارشات و مرجوعات، یکپارچه سازی صورت بگیرد.

میان صفحه سفارشات و مرجوعات، یک ویژگی مشترک يا join وجود دارد و برای یکپارچه سازی میان آن ها از عملگر join در نرم افزار استفاده می شود(افزوءه شدن ویژگی بازگشت با مقادیر بله و خیر). برای این کار، پیش از ورود داده های دو صفحه در عملگر join، شناسه سفارش را در هر کدام به عنوان دادهی شناسه يا ID مشخص کرده تا یکپارچه سازی بر مبنای یک مشخصه مشترک انجام شود. سپس، در ویژگی برگشت(Returned) در مقابل سفارش هایی که برگشت نخورده اند مقادیر خالی قرار می گیرد که با علامت سوال مشخص شده اند. بنابراین لازم است پس از یکپارچه سازی، مجدد رویکردی در برابر داده های مفقود داشته باشیم. این رکوردها با کلمهی No جایگزین خواهند شد.

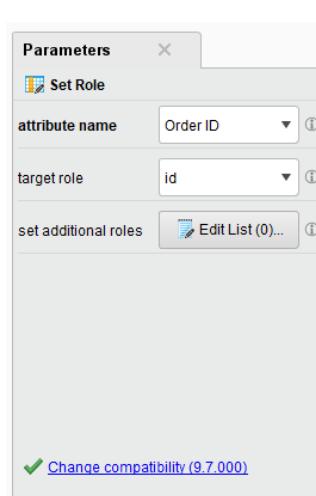


شکل ۱۷ – مدل یکپارچه سازی داده ها

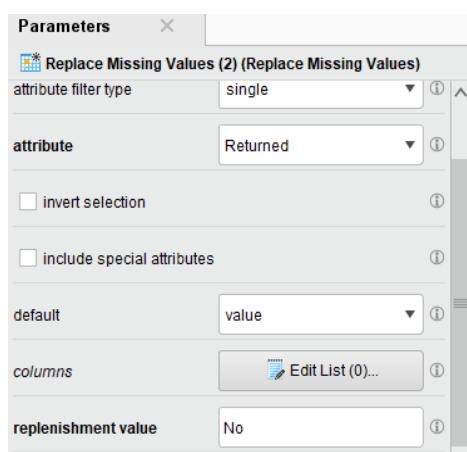
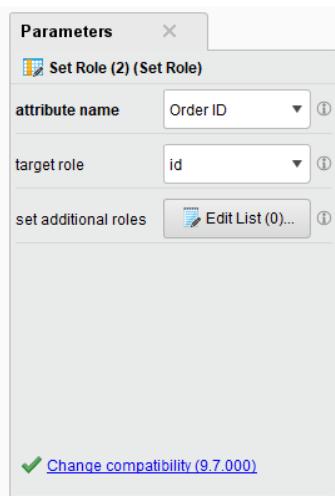
عملگرهایی که مستقیماً مربوط به یکپارچه سازی می باشند، با رسم مربع به دور آن ها مشخص شده اند. بقیه عملگر ها مربوط به مراحل پیشین است.



شکل ۱۹ – جزئیات عملگر join



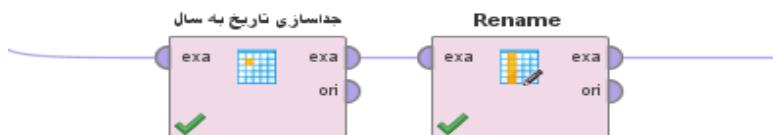
شکل ۱۸ – جزئیات دو عملگر تعیین نقش (Set Role)



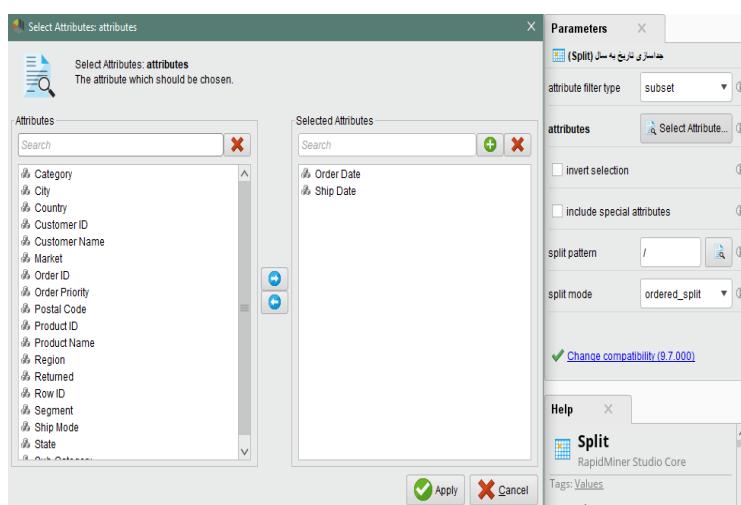
شکل ۲۰ – جزئیات عملگر جایگزینی مقادیر مفقود

شکل ۲۰ جزئیات افزودن مقادیر No در برابر سفارش هایی است که بازگشت ندارند.

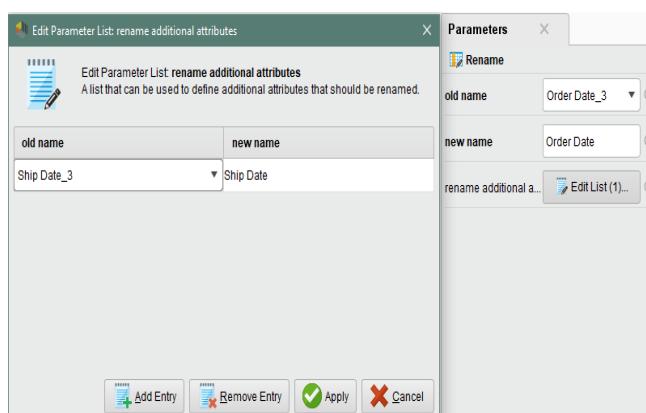
برای تبدیل داده، لازم است مقادیر درون ویژگی های مربوط به تاریخ دریافت و ارسال با استفاده از عملگر Split (جداسازی) تاریخ به سال) جداسازی شده و ویژگی های جدیدی که تنها تاریخ دریافت و ارسال سفارش را بر حسب سال نشان می دهند داشته باشیم. سپس با استفاده از عملگر Rename، نام جدیدی برای این ویژگی های جدید انتخاب می کنیم. شکل های زیر، جایگاه این عملگرهای در مدل و جزئیات آن ها را به خوبی نمایش می دهند.



شکل ۲۱ – عملگر های Split و Rename

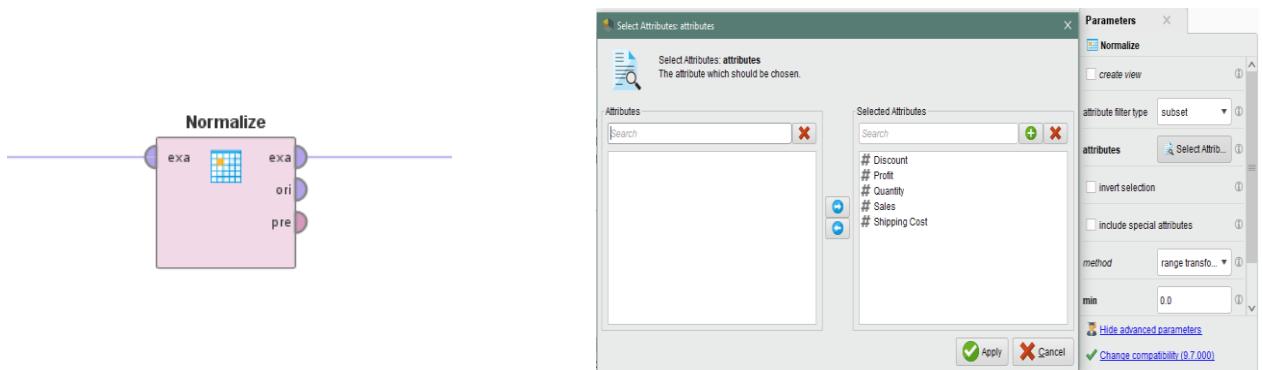


شکل ۲۲ – جزئیات عملگر Split



شکل ۲۳ – جزئیات عملگر Rename

همچنین، لازم است عملیات نرمال سازی بر روی ویژگی های کمی صورت بگیرد. نرمال سازی به افزایش دقت در خوش بندی کمک بسیاری خواهد کرد. برای این منظور، از عملگر Normalize و نرمال سازی در محدوده ۰ تا ۱ استفاده می کنیم.



شکل ۲۴ – مدل نرمال سازی

۱-۲-۳-۴- کاهش داده

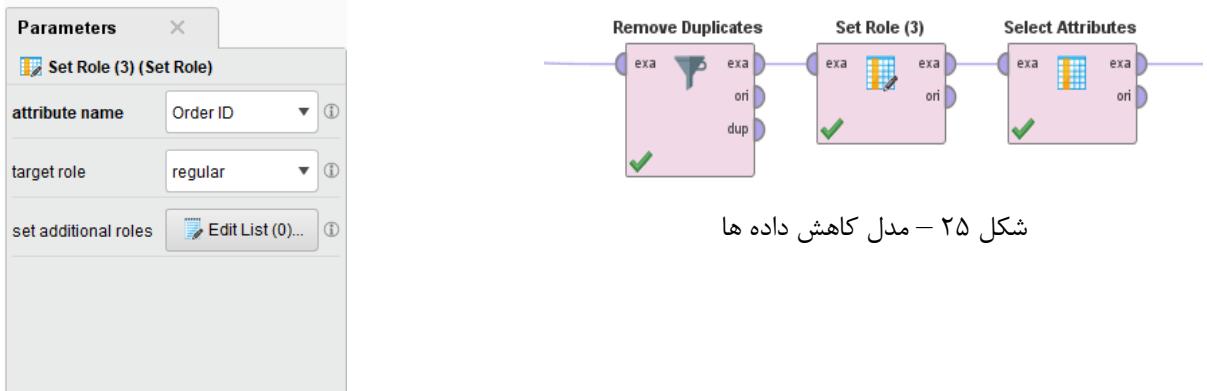
در این گام، لازم است مهم ترین ویژگی ها را انتخاب کرده و عملیات کاهش ابعاد داده ها صورت بگیرد(انتخاب زیر مجموعه ای از مشخصه ها و کاهش ستونی). مشخص است که برخی از ویژگی ها اهمیتی در فرآیند خوش بندی نخواهند داشت بنابراین از آن ها صرف نظر خواهد شد. این ویژگی ها عبارتند از:

مشخصه یا ID سطر، مشخصه یا ID سفارش، مشخصه یا ID مشتری، کد پستی، مشخصه یا ID محصول، نام مشتری و نام محصول. دیگر ویژگی ها اهمیتی به مراتب بیشتر از موارد مذکور دارند و از آن جا که نمی توان برای یک شهر و کشور به صورت جداگانه برنامه ریزی کرد و معمولاً دیدگاه مدیریتی به صورت کل نگرانه تراست(توجه بیشتر به ویژگی های بازار و منطقه)، پس شهر، ایالت و کشور را نیز حذف می کنیم.

با کاهش ۱۰ مورد از ابعاد داده ها، تعداد ابعاد یا ویژگی های باقی مانده که برای خوش بندی مورد استفاده قرار خواهند گرفت به ۱۵ خواهد رسید. ویژگی های نرمال شده جدیدی که ساختیم، به جای ویژگی های اصلی به کار خواهند رفت. ویژگی های حاصل عبارتند از:

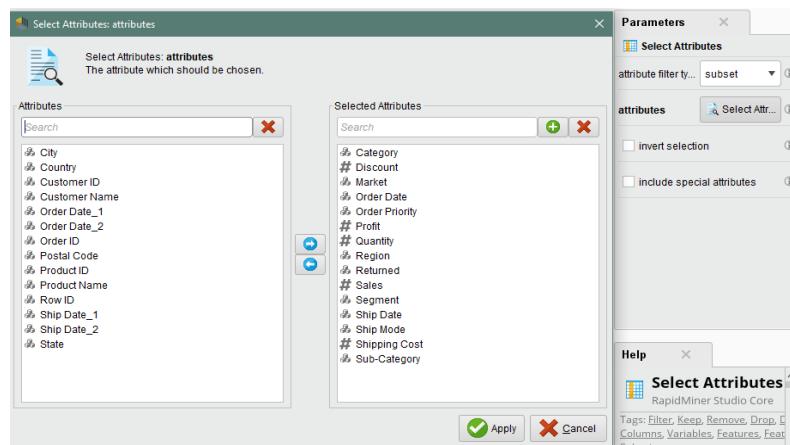
نوع ارسال، بخش یا Segment ، منطقه(برای مثال اقیانوسیه، آفریقای غربی و...)، بازار، دسته بندی یا مجموعه محصول، زیرمجموعه محصول، فروش نرمال شده، تعداد، تخفیف، سود نرمال شده ، هزینه ارسال نرمال شده و اولویت ارسال، بازگشت، تاریخ سفارش و تاریخ ارسال.

پیش از شروع کار، با استفاده از عملگر Remove Duplicates داده های تکراری را حذف کرده و سپس ادامه می دهیم. کاهش داده با استفاده از عملگر Select Attributes صورت می گیرد اما پیش از آن، بازهم باید از عملگر دیگری استفاده شود. این عملگر، عملگر Set Role است. پیشتر دیدیم که با استفاده از همین عملگر، ویژگی شناسه یا ID سفارش را به عنوان شناسه هی دو پایگاه داده ورودی سفارش و برگشت تنظیم کردیم join بتواند یکپارچه سازی را انجام دهد اما اگر این مشخصه هم چنان از نوع ID باشد، هرچند که در عملگر Select Attributes آن را انتخاب نکنیم، حذف نخواهد شد بنابراین این بار پیش از کاهش داده ها، نوع آن را به Regular تغییر می دهیم.



شکل ۲۵ – مدل کاهش داده ها

شکل ۲۶ – تغییر نوع ویژگی مشخصه سفارش (Order ID)



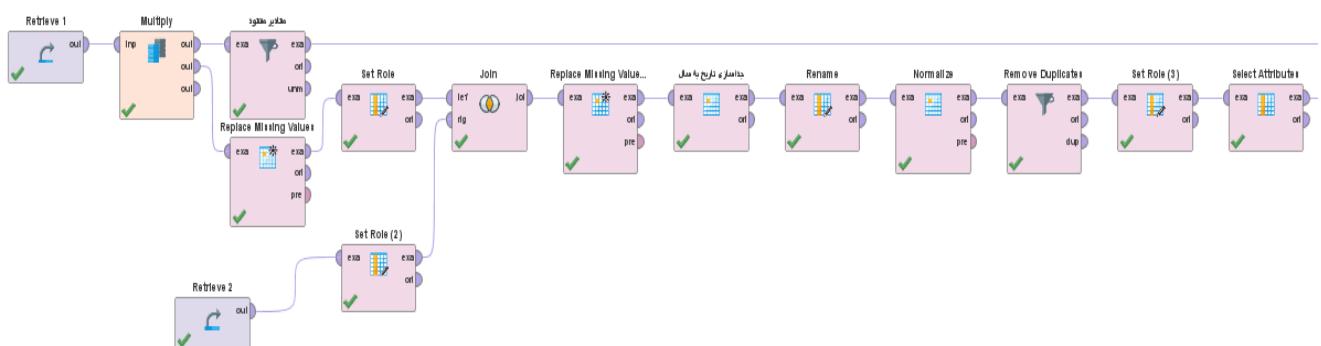
شکل ۲۷ – جزئیات عملگر کاهش ابعاد داده

۱-۲-۳-۵- گسسته سازی داده

به گسسته سازی نیازی نمی باشد زیرا داده های ما مناسب ایجاد خوشة می باشند. برای مثال اگر قصد ما خوشه بندی بر حسب فروش و منطقه باشد، گسسته سازی هیچ جایی نخواهد داشت.

۱-۲-۴- مدلسازی

مدل های داده کاوی در بخش ۲ و ۳ این پژوهش به تفصیل نشان داده خواهند شد اما مدل آماده سازی داده ها به صورت زیر می باشد:



شکل ۲۸ – مدل آماده سازی داده ها

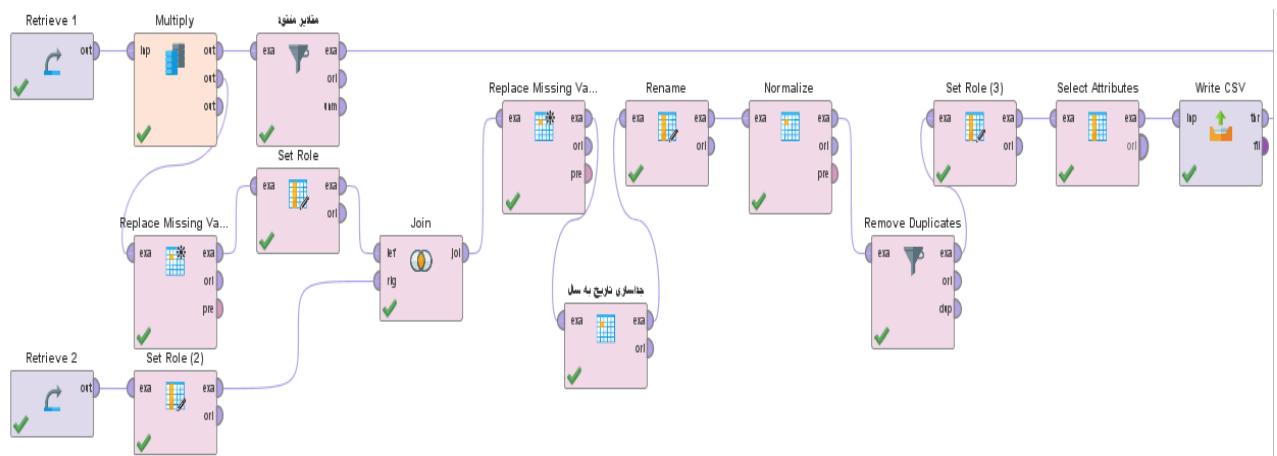
۱-۲-۵- ارزیابی

هر کدام از مدل های داده کاوی، پس از اجرا ارزیابی خواهد شد که مناسب بودن آن ارزیابی ها به نحوی نشان از آماده سازی مناسب داده ها نیز دارد.

۱-۲-۶- اجرا

مدل های داده کاوی در بخش ۲ و ۳ اجرا خواهند شد.

مدل آماده سازی داده ها نیز با استفاده از عملگر **Write CSV** نهایی خواهد شد و خروجی به صورت یک فایل با پسوند **CSV** و نام **Data Warehouse** ذخیره سازی خواهد شد. این فایل، انبار داده آماده برای داده کاوی می باشد.



شکل ۲۹ - مدل نهایی گام ۱

۱-۳-۱- تمايل به خوشه بندی

هر چند که لازم است این عملیات هرچه زودتر صورت بگیرد اما تا زمانی که یک پایگاه داده آماده و پاکسازی شده نداشته باشیم، امکان خطا وجود دارد و اکنون که پایگاه داده آماده است، برای بررسی تمايل داده ها به خوشه بندی، از روش هیستوگرام مسافت های جزئی استفاده می کنیم.

برای این منظور، لازم است که به صورت دو به دو، فاصله میان تمامی رکوردها را محاسبه کرده و سپس از مجموعه اعداد به دست آمده (فواصل دو به دو)، هیستوگرام نصب کرد. در پایگاه داده مورد استفاده، ۳ ویژگی عددی مناسب برای این کار وجود دارد:

۱- فروش نرمال شده

۲- سود نرمال شده

۳- هزینه ارسال نرمال شده

در نتیجه می توان اشیایی با ۳ مشخصه ساخت که مشخصه اول فروش نرمال شده، مشخصه دوم سود نرمال شده و مشخصه سوم هزینه ارسال نرمال شده را نشان دهد. این کار همانند حذف کردن همه ویژگی ها به جز ۳ مورد مذکور است. برای محاسبات از زبان برنامه نویسی پایتون کمک می گیریم (فایل کدها در پیوست) اما از آن جا که تعداد محاسبات بسیار زیاد خواهد

شد و نیاز به زمان زیاد یا سیستم قدرتمندی است، اشیا را با ۲ ویژگی در نظر می‌گیریم به این ترتیب ۳ نوع ترکیب برای هر شیء خواهیم داشت: زوج مرتب ۱ که به صورت (فروش، سود) است. زوج مرتب ۲ که به صورت (فروش، هزینه) است. زوج مرتب ۳ که به صورت (هزینه و سود) است. فاصله اقلیدسی هر کدام از این دو تایی های مرتب محاسبه و هیستوگرام آن ها در شکل های زیر نشان داده شده است.

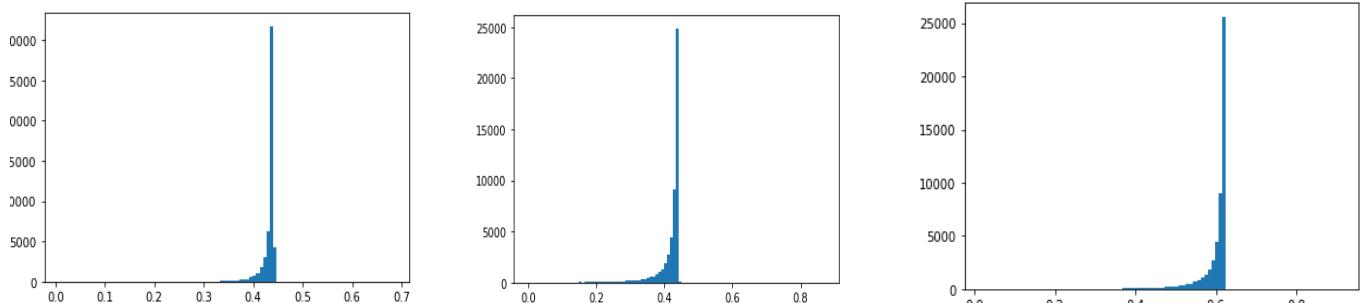
```

1 a=[]
2 b=[]
3 c=[]
4 import xlrd
5 loc = ("Edited_global_superstore_2016.xlsx")
6 wb = xlrd.open_workbook(loc)
7 sheet = wb.sheet_by_index(0)
8 sheet.cell_value(0, 0)
9
10 for i in range(1,51291):
11     a.append(sheet.cell_value(i, 10))
12 for j in range(1,51291):
13     b.append(sheet.cell_value(j, 13))
14 for k in range(1,51291):
15     c.append(sheet.cell_value(k, 14))
16
17 import numpy as np
18 t1=list(zip(a,b))
19 t2=list(zip(a,c))
20 #t3=list(zip(b,c))
21 t1=np.array(t1)
22 t2=np.array(t2)
23 #t3=np.array(t3)
24 d1=np.linalg.norm(t1-t2, axis=1)
25 #d2=np.linalg.norm(t1-t3, axis=1)
26 #d3=np.linalg.norm(t2-t3, axis=1)
27
28 from matplotlib import pyplot as plt
29 hist,bins=np.histogram(d1,100)
30 plt.hist(bins[:-1],bins,weights=hist)
31 plt.show()

```

شکل ۳۰ – سورس کد محاسبه فواصل اقلیدسی

نکته قابل توجه این است که کامنت می‌باشند (پیش از آن ها علامت # گذاشته شده است) هر بار به صورت جداگانه اجرا شده اند تا سرعت اجرا بالاتر رفته و سپس نمودار های حاصل در زیر رسم شده اند (به جای d1 در سطر ۲۹، یک بار d2 و یکبار d3 قرار داده شده است). فایل سورس کد پایتون، در پیوست های این پژوهش قرار دارد.

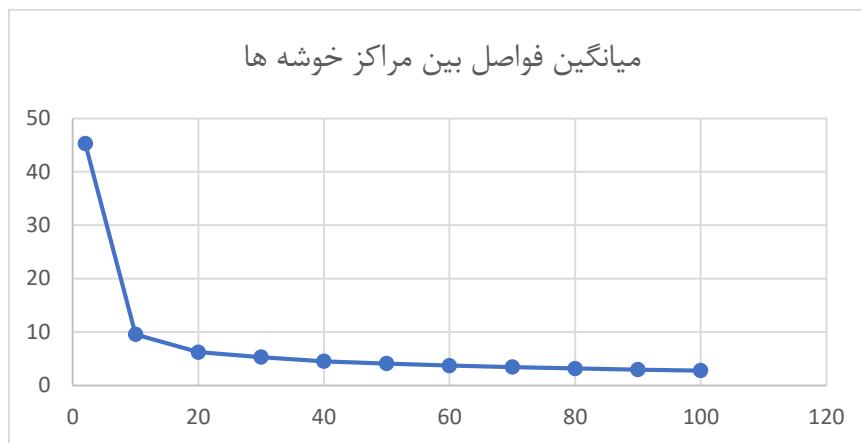


شکل ۳۱ – هیستوگرام فواصل زوج مرتب های ۱ و ۲ و ۳
شکل ۳۲ – هیستوگرام فواصل زوج مرتب های ۱ و ۲

وجود قله نشان از این دارد که میان داده ها نوعی همبستگی و در نتیجه تمایل به خوشه بندی وجود دارد. این قله ها در هر ۳ نمودار به چشم می خورد.

۴-۱- تعداد بهینه k

از روش ELBOW برای تعیین تعداد بهینه خوشه ها یا k استفاده شده است. در این روش خوشه بندی را با تعداد خوشه های مختلف انجام می‌شود و سپس میانگین فواصل درون خوشه ای بدست خواهد آمد(با استفاده از عملگر Performance) و در نهایت از اعداد به دست آمده، یک نمودار رسم خواهیم کرد. در هر نقطه ای از نمودار که شکست رخ دهد، تعداد بهینه خوشه ها یا k بدست آمده است(زیرا نشان می‌دهد کاهش فواصل درون خوشه ای و شاخص های ارزیابی خوشه ها پس از این نقطه چندان قابل توجه نیست).



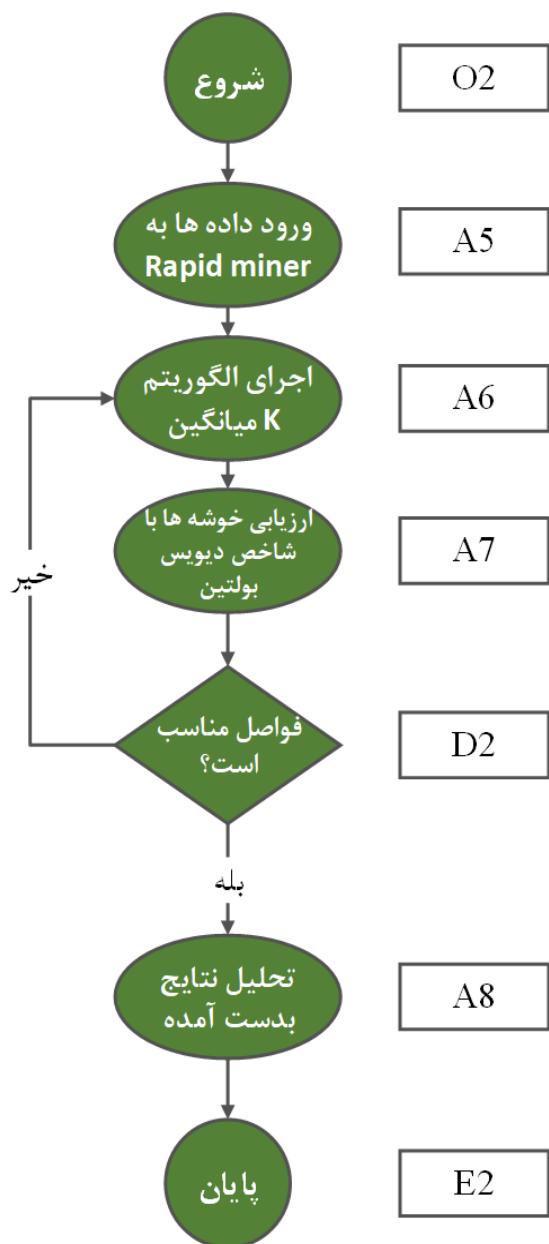
شکل ۳۴ - نمودار میانگین فواصل مرکز خوشه ها

همانطور که از شکل بالا مشخص است، شکست در نقطه ۱۰ رخ داده است که نشان می‌دهد تعداد بهینه خوشه ها عدد ۱۰ می‌باشد. این میانگین ها، در فایل اکسلی به نام Centroids نوشته شده و سپس نمودار آن ها رسم شده است(این فایل، در پیوست های پژوهش قرار دارد). وجود این شکست در نمودار نشان می‌دهد که افزایش تعداد خوشه ها تنها به میزان کمی باعث بهبود در خوشه بندی خواهد شد.

۲- گام دوم: خوشه بندی با استفاده از K-میانگین

پس از آماده سازی داده ها و مشخص شدن تعداد بهینه خوشه ها، عملیات خوشه بندی با استفاده از روش K-میانگین صورت می گیرد و سپس در صورت مناسب بودن خوشه بندی، نتایج حاصل تحلیل می شود.

فلوچارت این گام در شکل ۳۵ نشان داده شده است.



شکل ۳۵ - فلوچارت خوشه بندی با K-میانگین

۱-۲- ساخت و اجرای مدل

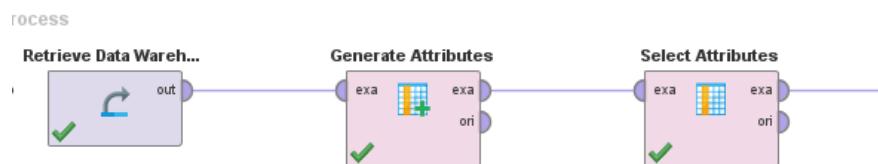
مدل را به چند روش ساخته و سپس نتایج به دست آمده با یکدیگر مقایسه شده اند.

در ابتدا، داده های حاصل از گام پیش که در واقع انبار داده می باشند، فراخوانی شده و وارد مدل می شوند. از آن جا که الگوریتم خوشه بندی، فاصله را بر مبنای عدد تعیین می کند، بنابراین به هر داده از ویژگی های کیفی یک عدد نسبت داده می شود. عملگر Generate Attributes این کار را میسر می کند. ویژگی های جدید با قرار دادن واژه New پیش از نام اصلی آن ها ساخته می شوند (برای مثال ویژگی New Ship Mode، جایگزین ویژگی Ship Mode خواهد شد). در ادامه، از عملگر Select Attribute برای انتخاب ویژگی های ساخته شده جدید و ویژگی های کمی استفاده می کنیم.

روش تخصیص عدد به متغیرهای کیفی به دو صورت است. برای متغیرهای کیفی رتبه ای، از فرمول زیر برای نرمال سازی و تخصیص برچسب عددی استفاده می شود. در این فرمول، r نشان دهنده رتبه ای کیفی (برای مثال به ازای بیشترین اولویت ارسال یا critical که رتبه ۱ را دارد $r = 1$) و p نشان دهنده کل حالات ممکن است (کل رتبه های کیفی).

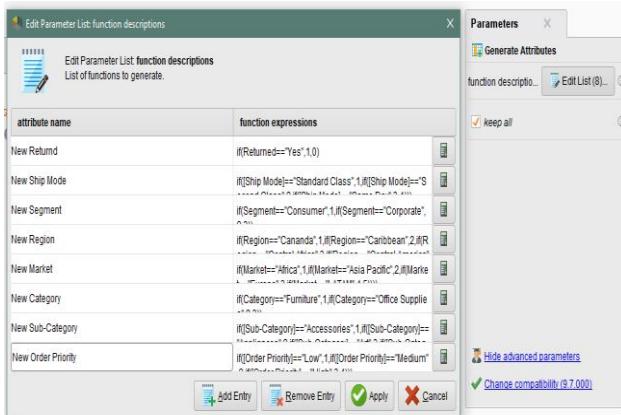
$$z = \frac{r - 1}{p - 1}$$

برای متغیرهای کیفی اسمی نیز می توان برچسب عددی تخصیص داد اما از آنجا که الگوریتم خوشه بندی k -میانگین اساساً برای متغیرهای کمی است، چنین رویکردی توصیه نمی شود.^{۱۰} برای متغیر باینری بازگشت (Returned)، در صورت "yes" یا بله بودن عدد ۱ و در غیر این صورت عدد ۰ تخصیص داده شده است. بنابراین علاوه بر ویژگی های کمی موجود، از ۳ ویژگی کیفی با برچسب های عددی (کمی شده) استفاده خواهیم کرد که عبارتند از بازگشت، نوع ارسال و اولویت ارسال. در اولین نوع مدلسازی، تنها از ویژگی ها کمی استفاده خواهیم کرد. سپس ویژگی های کیفی ترتیبی و سپس ویژگی های کیفی اسمی نیز اضافه می شوند.



شکل ۳۶ – سه مرحله ابتدایی گام ۲

^{۱۰} متغیرهای کیفی موجود در انبار داده بی استفاده نمی باشند بلکه رویکرد خوشه بندی به طور کامل مناسب استفاده از چنین متغیرهایی نمی باشد اما این موضوع به این معنا نیست که این ویژگی ها زاید می باشند زیرا در قسمت ۱ این پژوهش (آماده سازی داده ها)، ویژگی های کم اهمیت حذف شدند. در ادامه رویکرد برچسب گذاری به این متغیرها انجام می شود و نتایج بدست آمده با نتایج قبلی مقایسه خواهند شد. هرچند که به وضوح نتایج بدون این کار بهتر است.



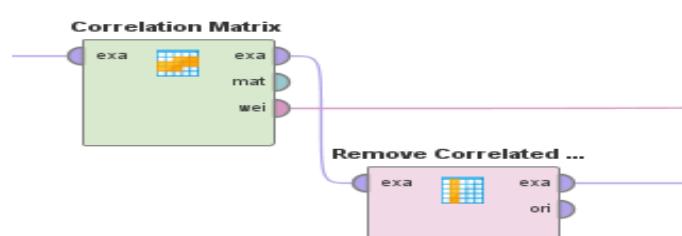
شكل ۳۸ – جزئیات عملگر Generate Attributes

Format your columns.							
	Date format	Enter value...					
1	Sales	real	Discount	real	Profit	real	Shipping C...
2	-0.051	-0.673	0.192	0.249	No	First Class	Second Class
3	7.098	-0.203	-1.819	15.660	No	First Class	First Class
4	10.102	-0.203	5.109	15.518	No	First Class	First Class
5	5.423	-0.203	-0.717	15.425	No	Same Day	Standard Class
6	5.301	-0.673	1.621	15.301	No	Second Class	First Class
7	5.362	-0.203	4.211	15.201	No	Standard Class	First Class
8	3.229	-0.673	3.073	15.156	No	First Class	First Class
9	10.245	-0.673	5.547	14.870	No	First Class	First Class
10	0.195	-0.673	0.149	-0.022	No	First Class	First Class
11	-0.406	0.268	-0.133	-0.269	No	First Class	First Class
	-0.469	-0.673	-0.137	-0.368	No	First Class	First Class

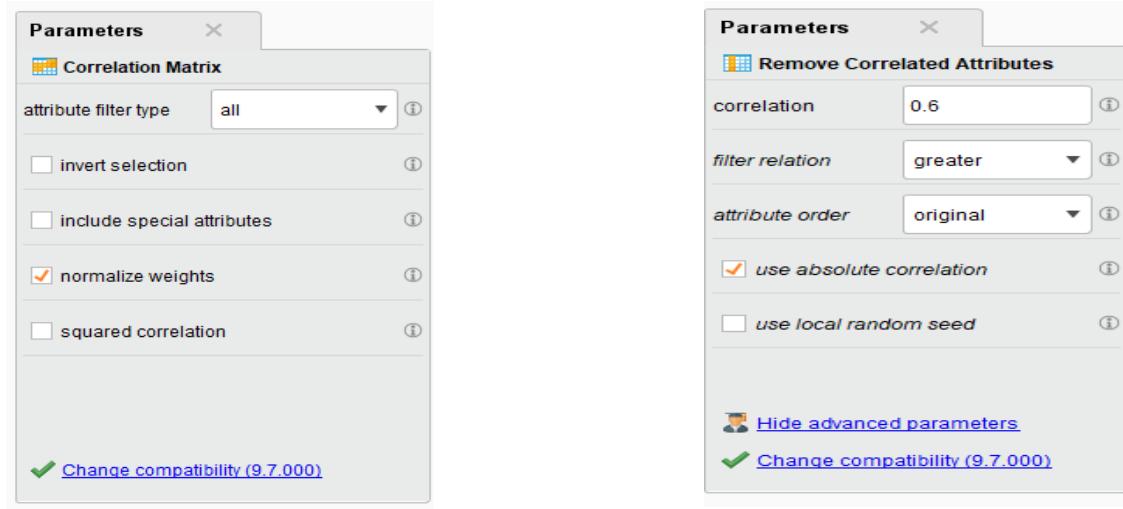
شكل ۳۷ – وارد کردن انباره داده

شكل ۳۹ – جزئیات عملگر انتخاب ویژگی

الگوریتم خوش بندی، برای اجرای مناسب نیاز دارد همبستگی میان ویژگی داده ها زیاد نباشد بنابراین در ادامهی کار، با استفاده از عملگر Remove Correlated، ویژگی های همبسته شناسایی و با استفاده از عملگر Correlation Matrix، مشخصه هایی که میزان همبستگی میان آن ها بیش از ۶۰٪ باشد حذف می شود و در گام آخر، الگوریتم خوش بندی k-میانگین اجرا می شود.

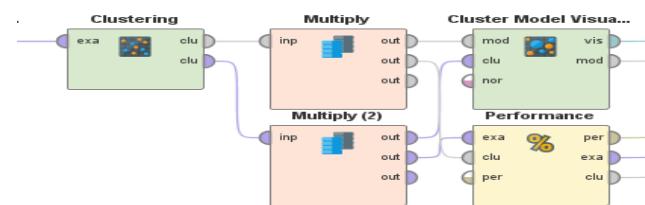


شكل ۴۰ – عملگرهای تشخیص و حذف مشخصه های همبسته

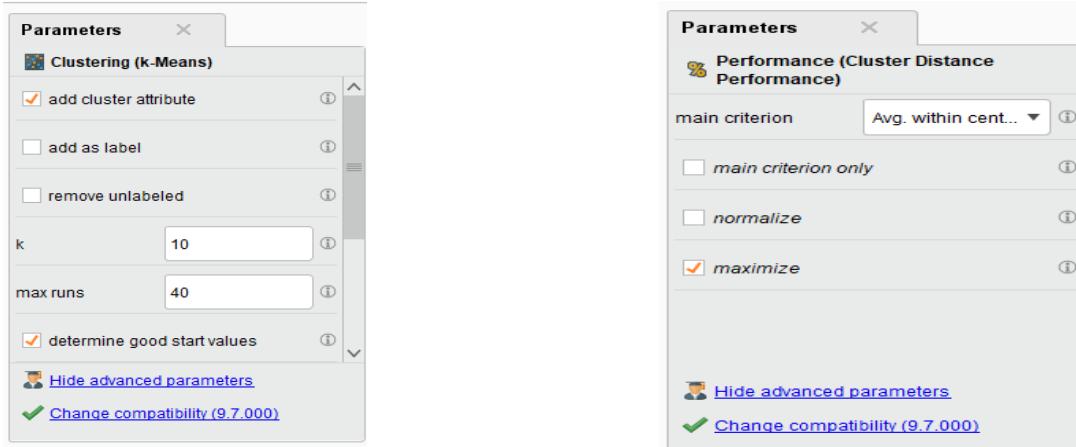


شکل ۴۱ – جزئیات عملگرهای تشخیص و حذف مشخصه های همبسته

در انتهای، می توان الگوریتم خوش بندی با تکنیک K-میانگین را اجرا کرد. پس از اجرا، لازم است مدل اعتبار سنجی شود بنابراین پس از آن، از عملگر **Performance** استفاده می شود. همچنین از عملگر **Cluster Model Visualizer** برای مصور سازی بهتر استفاده خواهد شد. از آنجا که لازم است ۴ خروجی از عملگر **Clustering** دریافت شود(برای هر یک از ۲ عملگر پس از آن دو خروجی)، از ۲ عملگر **Multiply** استفاده خواهد شد. شکل های زیر، گام های انتهاهی مدل را به همراه جزئیات نشان می دهند.

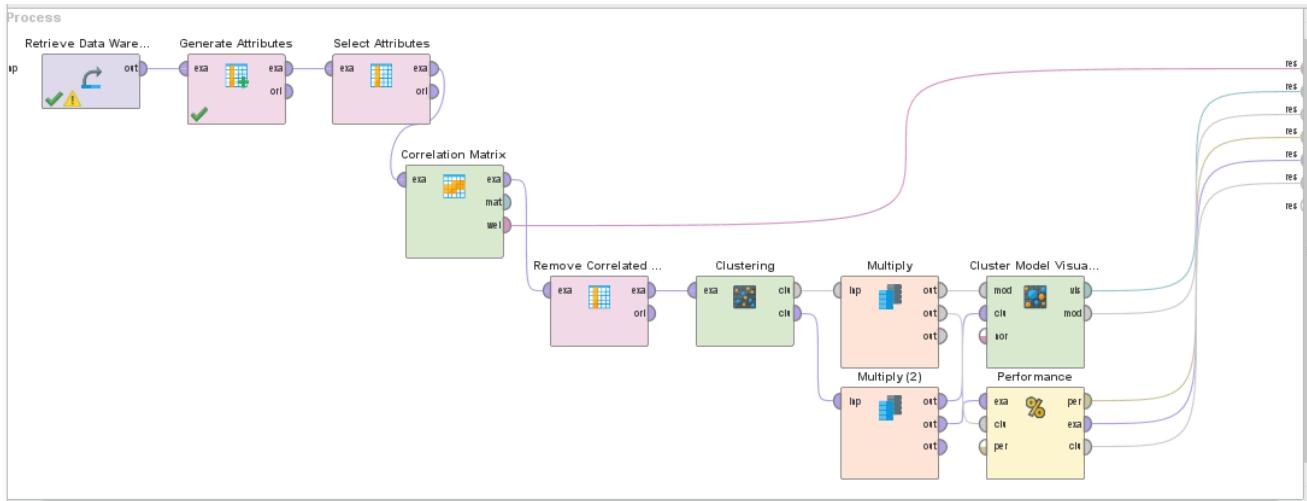


شکل ۴۲ – عملگر های خوش بندی و سنجش اعتبار



شکل ۴۳ – جزئیات عملگرهای خوش بندی و اعتبار سنجی

در شکل زیر، تصویر کلی مدل نهایی برای خوش بندی با استفاده از الگوریتم k -میانگین، نشان داده شده است.



شکل ۴۴ - مدل k -میانگین

۲-۲- تحلیل نتایج به دست آمده

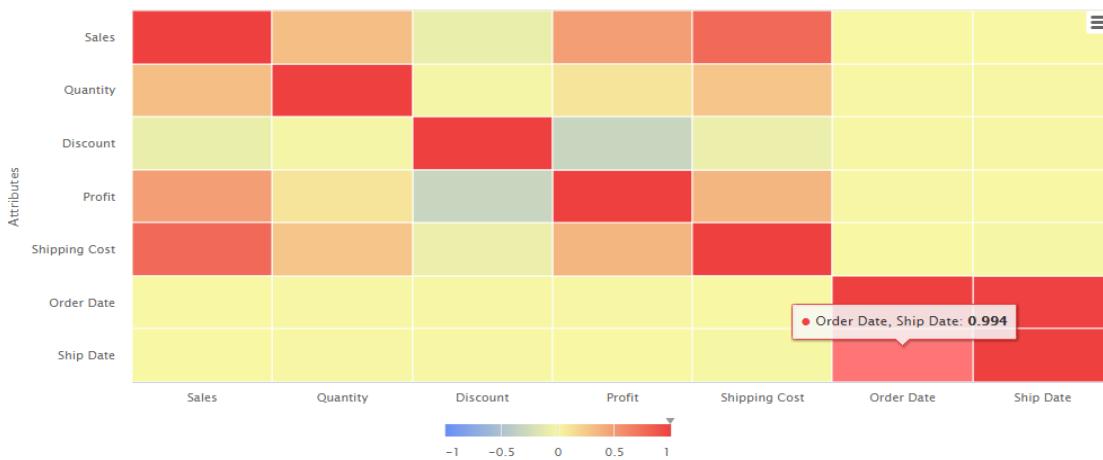
شاخص ارزیابی (دیویس بولدین) برابر با 0.704 ± 0.032 می باشد که مقدار مناسبی است. همچنین، متوسط فواصل درون خوشه ای نیز 0.032 ± 0.003 می باشد. بنابراین، مدل از اعتبار خوبی برخوردار است و نیاز به اجرای مجدد و تغییرات نیست.

دو ویژگی تاریخ ارسال و هزینه ارسال به دلیل همبستگی بالا با ویژگی های تاریخ سفارش و فروش حذف شده اند.

Attribut...	Sales	Quantity	Discount	Profit	Shipping Cost	Order D...	Ship Date
Sales	1	0.314	-0.087	0.485	0.768	-0.003	-0.003
Quantity	0.314	1	-0.020	0.104	0.272	-0.005	-0.005
Discount	-0.087	-0.020	1	-0.317	-0.078	-0.006	-0.006
Profit	0.485	0.104	-0.317	1	0.354	0.003	0.002
Shipping...	0.768	0.272	-0.078	0.354	1	-0.003	-0.004
Order Da...	-0.003	-0.005	-0.006	0.003	-0.003	1	0.994
Ship Date	-0.003	-0.005	-0.006	0.002	-0.004	0.994	1

شکل ۴۵ - جدول همبستگی ها

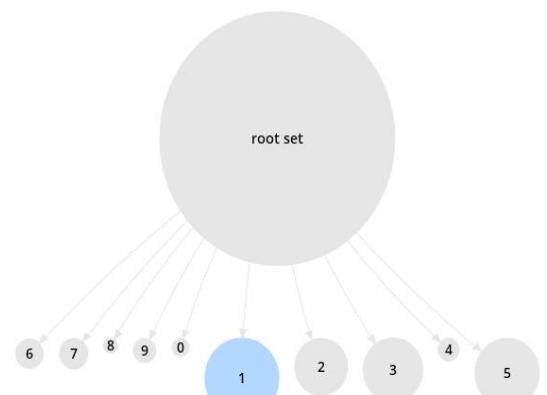
نمودار زیر نیز میزان همبستگی بین مشخصه ها را به صورت گرافیکی نمایش داده است. افزایش تراکم رنگ قرمز به معنای همبستگی بیشتر است. برای مثال، میزان همبستگی بین تاریخ سفارش و تاریخ ارسال 0.994 ± 0.006 است که به دلیل بزرگتر بودن از آستانه‌ی تعریف شده در مدل (0.6) حذف شده است.



شکل ۴۶ – نمودار همبستگی بین مشخصه ها

تعداد رکوردهای هر خوش به این صورت است:

Cluster 0: 2236 items	Cluster 7: 3761 items
Cluster 1: 10095 items	Cluster 8: 1901 items
Cluster 2: 7080 items	Cluster 9: 2986 items
Cluster 3: 7990 items	Total number of items: 51212
Cluster 4: 2808 items	
Cluster 5: 8706 items	
Cluster 6: 3649 items	



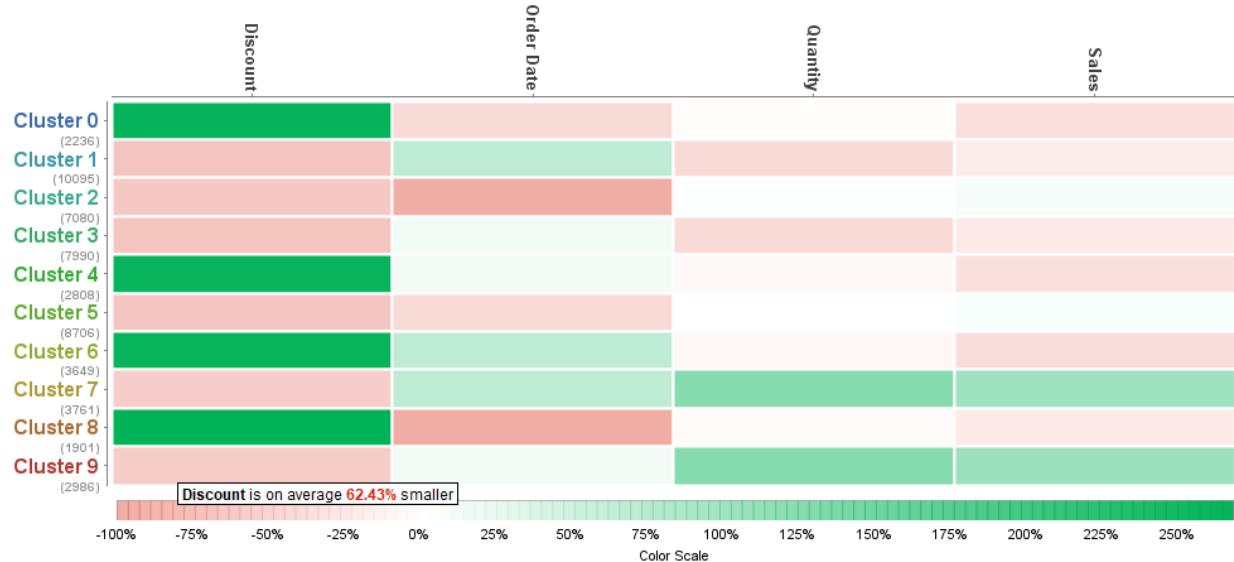
شکل ۴۷ – مصور سازی تعداد اعضای هر خوش

خوشی ۱ بیشترین تعداد رکوردها را شامل می شود و خوش ۸ کمترین تعداد. مصور سازی تعداد اعضا با استفاده از شکل زیر قابل رویت می باشد و بزرگترین خوش به رنگ آبی مشخص شده است. مراکز خوش ها نیز به صورت زیر است.

Cluster	Sales	Quantity	Discount	Profit	Order Date
Cluster 0	0.007	0.185	0.614	0.435	2013
Cluster 1	0.008	0.103	0.049	0.443	2015
Cluster 2	0.012	0.195	0.055	0.444	2012
Cluster 3	0.008	0.106	0.048	0.443	2014
Cluster 4	0.007	0.175	0.607	0.435	2014
Cluster 5	0.012	0.192	0.051	0.444	2013
Cluster 6	0.006	0.172	0.607	0.435	2015
Cluster 7	0.022	0.435	0.065	0.447	2015
Cluster 8	0.008	0.180	0.620	0.435	2012
Cluster 9	0.022	0.435	0.063	0.447	2014

شکل ۴۸ – مراکز خوش ها

روش جالبی که برای مصورسازی خوشه ها وجود دارد، نقشه حرارتی است. در این نقشه، به ازای مقادیر مشخصه های هر خوشه، یک مستطیل رنگی رسم شده است که هر چه رنگ آن به سبز تمایل داشته باشد نشان از بزرگتر بودن مرکز خوشه مورد نظر در آن ویژگی نسبت به سایر خوشه ها و هر چه رنگ آن قرمز تر باشد کوچکتر است. برای مثال در شکل زیر نشان داده شده است که متوسط تخفیف های خوشه ۹، به میزان ۶۲,۴۳٪ کوچکتر از متوسط میزان موجود است.



شکل ۴۹ - نقشه حرارتی خوشه ها و مشخصه ها

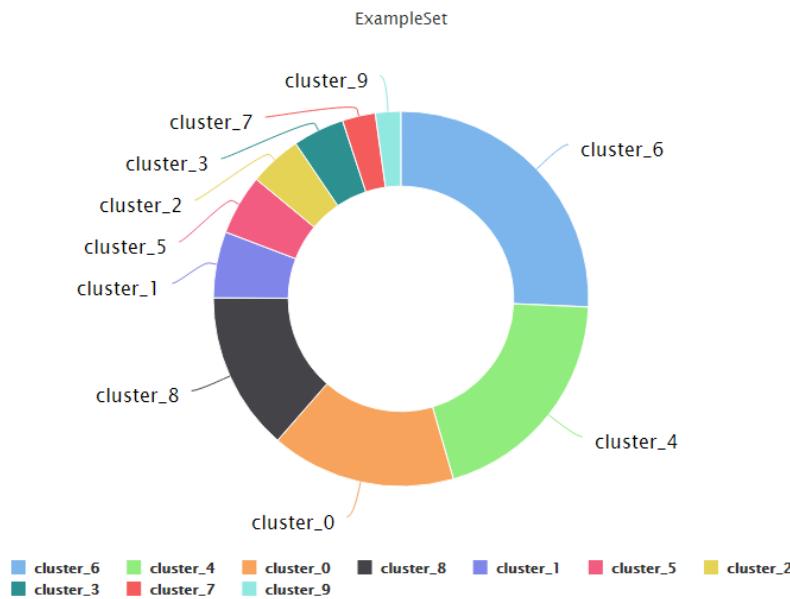
اطلاعات کلی خوشه ها در شکل زیر قابل مشاهده است.

Cluster 0	2,236	Average Distance: 0.055
Discount is on average 264.70% larger, Order Date is on average 43.74% smaller, Sales is on average 38.02% smaller		
Cluster 1	10,095	Average Distance: 0.014
Discount is on average 71.00% smaller, Order Date is on average 68.78% larger, Quantity is on average 45.72% smaller		
Cluster 2	7,080	Average Distance: 0.040
Order Date is on average 100.00% smaller, Discount is on average 67.37% smaller, Sales is on average 9.19% larger		
Cluster 3	7,990	Average Distance: 0.014
Discount is on average 71.49% smaller, Quantity is on average 44.36% smaller, Sales is on average 25.57% smaller		
Cluster 4	2,808	Average Distance: 0.047
Discount is on average 260.76% larger, Sales is on average 38.10% smaller, Order Date is on average 12.52% larger		

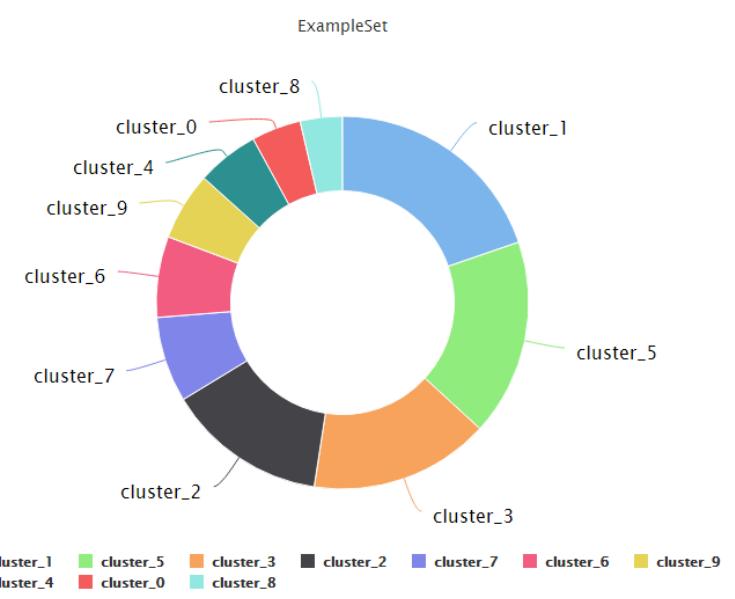
Cluster 5	8,706	Average Distance: 0.038
Discount is on average 69.43% smaller, Order Date is on average 43.74% smaller, Sales is on average 8.64% larger		
Cluster 6	3,649	Average Distance: 0.048
Discount is on average 260.80% larger, Order Date is on average 68.78% larger, Sales is on average 41.30% smaller		
Cluster 7	3,761	Average Distance: 0.035
Quantity is on average 128.61% larger, Sales is on average 101.75% larger, Order Date is on average 68.78% larger		
Cluster 8	1,901	Average Distance: 0.054
Discount is on average 268.42% larger, Order Date is on average 100.00% smaller, Sales is on average 25.28% smaller		
Cluster 9	2,986	Average Distance: 0.034
Quantity is on average 128.61% larger, Sales is on average 104.74% larger, Discount is on average 62.43% smaller		

شکل ۵۰ – خلاصه اطلاعات خوشه ها

اطلاعات آماری مناسبی نیز وجود دارند که برای نتیجه گیری نهایی مفید خواهند بود.

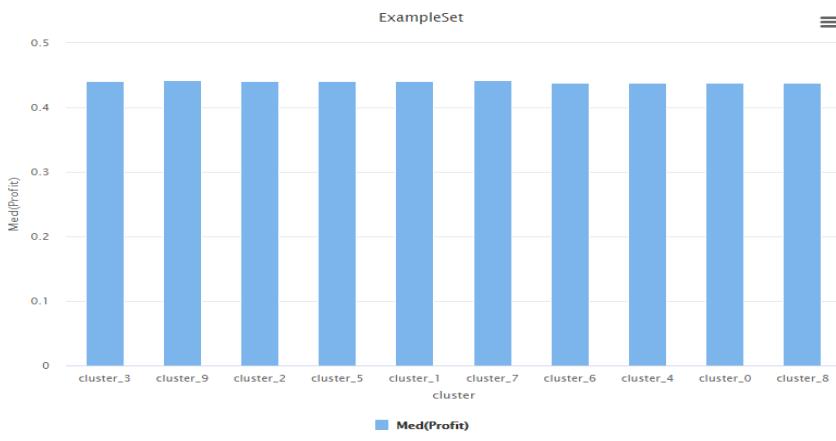


شکل ۵۲ – خوشها بر اساس مجموع تخفیف

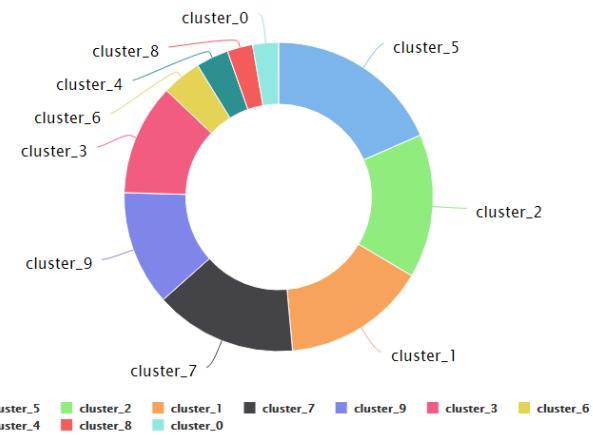


شکل ۵۱ – خوشها بر اساس مجموع سودآوری

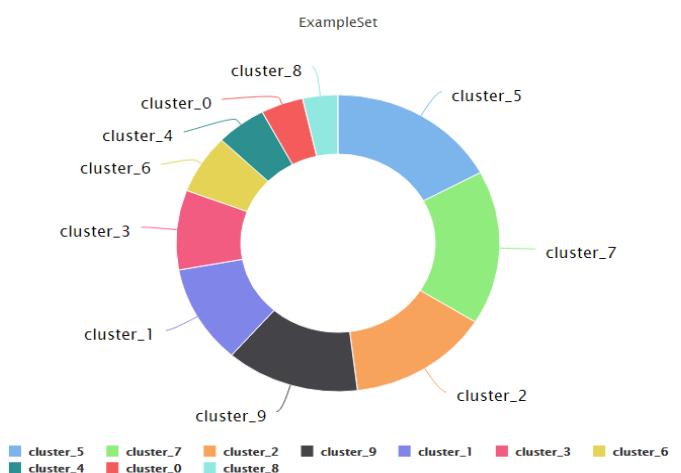
همانطور که در شکل های بالا می بینید، خوشه ۱ بیشترین و خوشه ۸ کمترین مجموع سودآوری را دارد. همچنین، خوشه ۶ بیشترین و خوشه ۹ کمترین مجموع تخفیف را دارد. همچنین می دانیم که خوشه ۱، بیشترین تعداد رکوردها را در بر می گیرد بنابراین آیا نتیجه گیری اینکه این خوشه واقعاً سودآورترین خوشه است نتیجه گیری مناسبی است؟ در جدول مرکز خوشها دیده شد که مقادیر مربوط به سود به ازای خوشه های مختلف بسیار نزدیک به هم می باشند بنابراین پاسخ این است که خوشه ۱ مانند دیگر خوشها از نظر متوسط سودآوری می باشد اما با این حال، بخش زیادی از رکوردها و در نتیجه مشتریان را شامل می شود. شکل صفحه بعد نیز این موضوع را نشان می دهد.



شکل ۵۳ - میانه سودآوری خوشه ها



شکل ۵۴ - خوشه ها بر اساس مجموع میزان درآمدی فروش



شکل ۵۵ - خوشه ها بر اساس مجموع تعداد فروش

خوشه ۵ بیشترین و خوشه ۸ کمترین تعداد فروش را دارد. خوشه ۵ بیشترین و خوشه ۰ کمترین مجموع درآمد حاصل از فروش را دارند. در جدول زیر، خلاصه از نمودار ها و اطلاعات وجود دارد. رتبه های درون جدول، بر اساس متوسط اعداد(رجوع شود به جدول مراکز خوشه ها) نوشته شده است.

جدول ۲ - رتبه بندی خوشه ها بر اساس ویژگی های مختلف

نام خوشه	رتبه سودآوری	رتبه تخفیف	رتبه تعداد فروش	رتبه درآمد فروش	رتبه تعداد تراکنش ها	میانگین سال
خوشه ۰	۷	۲	۵	۸	۹	۲۰۱۳
خوشه ۱	۵	۹	۱۰	۵	۱	۲۰۱۵
خوشه ۲	۳	۷	۳	۳	۴	۲۰۱۲
خوشه ۳	۶	۱۰	۹	۶	۳	۲۰۱۴
خوشه ۴	۸	۳	۷	۹	۸	۲۰۱۴
خوشه ۵	۴	۸	۴	۴	۲	۲۰۱۳
خوشه ۶	۹	۴	۸	۱۰	۶	۲۰۱۵
خوشه ۷	۱	۵	۱	۱	۵	۲۰۱۵
خوشه ۸	۱۰	۱	۶	۷	۱۰	۲۰۱۲
خوشه ۹	۲	۶	۲	۲	۷	۲۰۱۴

خوشه صفر: شامل ۴ درصد از تراکنش ها - خوشه ای پر تخفیف و کم فروش(فروش درآمدی) - سودآوری در رتبه ۷ - از نظر تعداد در میانه(رتبه ۵)

خوشه یک: شامل ۱۹ درصد تراکنش ها - بزرگترین بخش تراکنش ها - تخفیف و تعداد فروش پایین تر از متوسط - از نظر تعداد فروش کمترین رتبه و رتبه تخفیف ۹(میزان تخفیف ها کم) اما از نظر سودآوری و درآمد فروش در میانه(رتبه ۵)

خوشه دو: رتبه ۳ در سودآوری و درآمد و تعداد فروش - از نظر تخفیف در رتبه ۷ - شامل ۱۳ درصد تراکنش ها

خوشه سه: ۱۵ درصد تراکنش ها - تعداد فروش و میزان تخفیف کم - سومین خوشه بزرگ سودآوری متوسط - تخفیف کم - درآمد متوسط - تعداد فروش متوسط

خوشه چهار: شامل ۵ درصد تراکنش ها - تعداد فروش کمتر از متوسط اما تخفیف بالاتر از متوسط(رتبه ۳).

خوشه پنج: ۱۷ درصد تراکنش ها(رتبه ۲ از نظر تعداد) - تخفیف پایین تر از متوسط و فروش بالاتر از متوسط - خوشه ای سودآور(رتبه ۴)

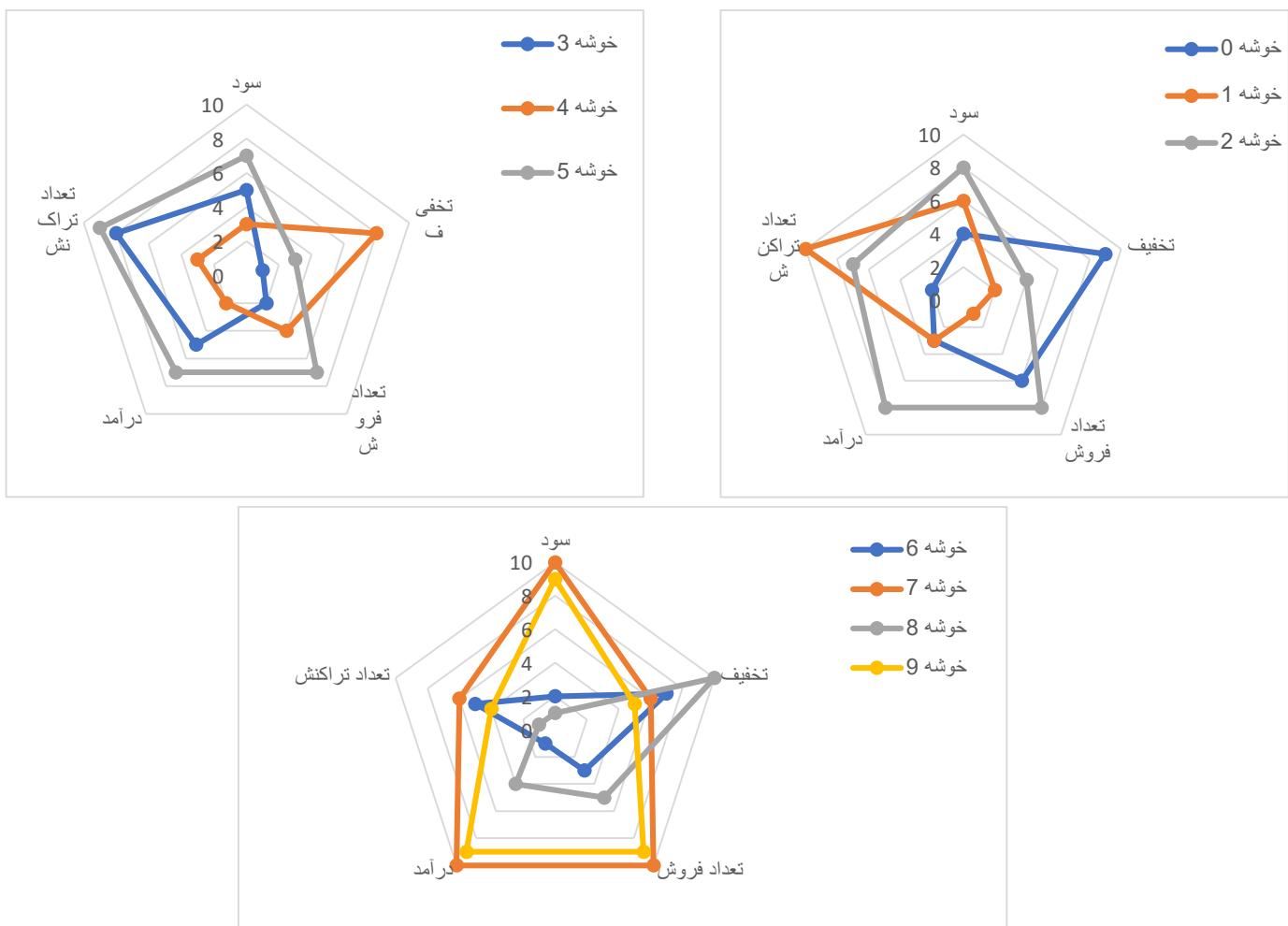
خوشه شش: ۷ درصد از تراکنش‌ها – تخفیف بالاتر و فروش پایین تر از متوسط – از کم سود ترین خوشه‌ها(رتبه ۹)

خوشه هفت: سودآورترین بخش مشتریان – رتبه ۶ در تخفیف – فروش و تعداد فروش رتبه ۲ بنابراین می‌توان خوشه‌ای پر فروش دانست؛ چه تعداد و چه درآمد – شامل ۷ درصد از تراکنش‌ها

خوشه هشت: کوچکترین خوشه(۳ درصد تراکنش‌ها) – سود و درآمد کمترین و تخفیف بیشترین میزان -

خوشه نه: خوشه‌ای پر فروش، پر درآمد و پر سود – تعداد و درآمد بالاتر و تخفیف پایین تر از متوسط – ۵ درصد تراکنش‌ها و خوشه‌ای کوچک(رتبه ۷)

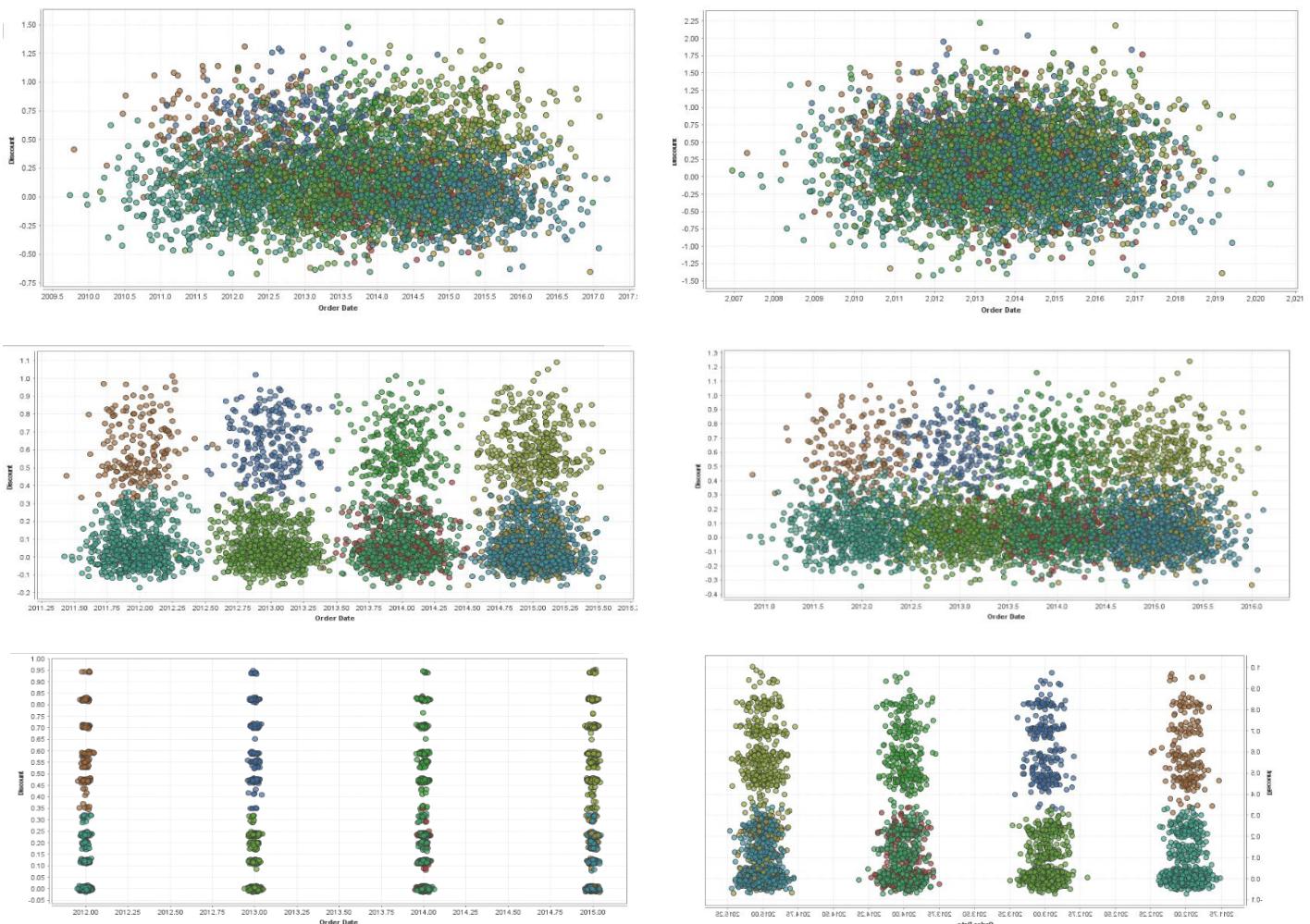
به نظر می‌رسد در بیشتر تراکنش‌های شرکت، افزایش در تعداد فروش باعث افزایش تخفیف(سرشکن کردن قیمت) نشده است اما همچنان این نوع تراکنش‌ها از نظر تعداد فروش رتبه‌های بالا را دارند. با همه این تفاسیر، سیاست فروش به گونه‌ای بوده که نفاوت زیادی میان سودآوری خوشه‌ها(بخش‌ها=segments) نبوده است.



شکل ۵۶ – نمایش خوشه‌ها بر اساس رتبه هر مشخصه^{۱۶}

^{۱۶} در این نمودارها برای جلوگیری از خطا اعداد مربوط به رتبه‌ها به صورت معکوس(رتبه ۱۰ = رتبه ۱) شماره گذاری و رسم شده است. برای مثال خوشه‌ای که بیشترین تخفیف را دارد رتبه ۱ را بدست می‌آورد اما در نمودار به مرکز چسبیده و تصور می‌شود در این خوشه نخفیف زیادی وجود ندارد در حالی که موضوع برعکس است.

در نهایت، شکل های زیر برای نشان دادن پراکندگی رکوردها در خوشه ها استفاده شده اند.



شکل ۵۷ – تخصیص رکوردها به هر خوشه

۳-۲ - مدل های دیگر

در ادامه، یکبار دیگر خوشه بندی را با دخیل کردن متغیرهای کیفی به صورت عددی شده انجام می‌دهیم. در اولین مرحله متغیر های کیفی رتبه ای نوع ارسال و اولویت ارسال دخیل می‌شوند. نتایج به صورت زیر است.

شاخص این مدل ۹۸۵ و میانگین فواصل بین مراکز خوشه ها ۱۳۶ است بنابراین اعتبار مدل کمتر شده است(اما قابل قبول).

Cluster 0: 8538 items

Cluster 4: 6868 items

Cluster 8: 2665 items

Cluster 1: 3478 items

Cluster 5: 11021 items

Cluster 9: 3006 items

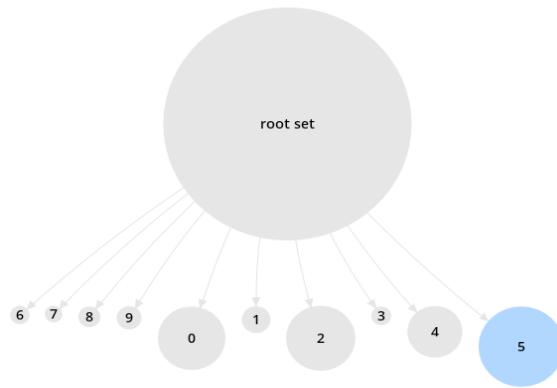
Cluster 2: 8774 items

Cluster 6: 2345 items

Total number of items: 51212

Cluster 3: 2404 items

Cluster 7: 2113 items



شکل ۵۸ – بزرگی خوشة ها

همبستگی میان مشخصه ها مطابق قبل می باشد.

Attribut...	Sales	Quantity	Discount	Profit	Shipping Cost	Order Date	Ship Date	New Ship Mode	New Order Priority
Sales	1	0.314	-0.087	0.485	0.768	-0.003	-0.003	0.001	-0.003
Quantity	0.314	1	-0.020	0.104	0.272	-0.005	-0.005	0.005	0.002
Discount	-0.087	-0.020	1	-0.317	-0.078	-0.006	-0.006	-0.010	0.005
Profit	0.485	0.104	-0.317	1	0.354	0.003	0.002	0.003	-0.002
Shipping...	0.768	0.272	-0.078	0.354	1	-0.003	-0.004	-0.145	-0.175
Order Da...	-0.003	-0.005	-0.006	0.003	-0.003	1	0.994	-0.002	0.014
Ship Date	-0.003	-0.005	-0.006	0.002	-0.004	0.994	1	0.001	0.018
New Shi...	0.001	0.005	-0.010	0.003	-0.145	-0.002	0.001	1	0.449
New Ord...	-0.003	0.002	0.005	-0.002	-0.175	0.014	0.018	0.449	1

شکل ۵۹ – جدول همبستگی میان مشخصه ها

Cluster	Sales	Quantity	Discount	Profit	Order Date	New Ship Mode	New Order Prior...
Cluster 0	0.011	0.189	0.164	0.442	2013	0.928	0.594
Cluster 1	0.011	0.188	0.166	0.442	2015	0.084	0.352
Cluster 2	0.012	0.194	0.050	0.444	2014	0.916	0.575
Cluster 3	0.011	0.195	0.176	0.442	2013	0.152	0.318
Cluster 4	0.011	0.195	0.173	0.442	2012	0.926	0.591
Cluster 5	0.012	0.192	0.050	0.444	2015	0.914	0.576
Cluster 6	0.007	0.181	0.606	0.435	2014	0.902	0.578
Cluster 7	0.011	0.181	0.181	0.442	2012	0.162	0.294
Cluster 8	0.011	0.191	0.156	0.442	2014	0.089	0.352
Cluster 9	0.007	0.180	0.607	0.435	2015	0.909	0.582

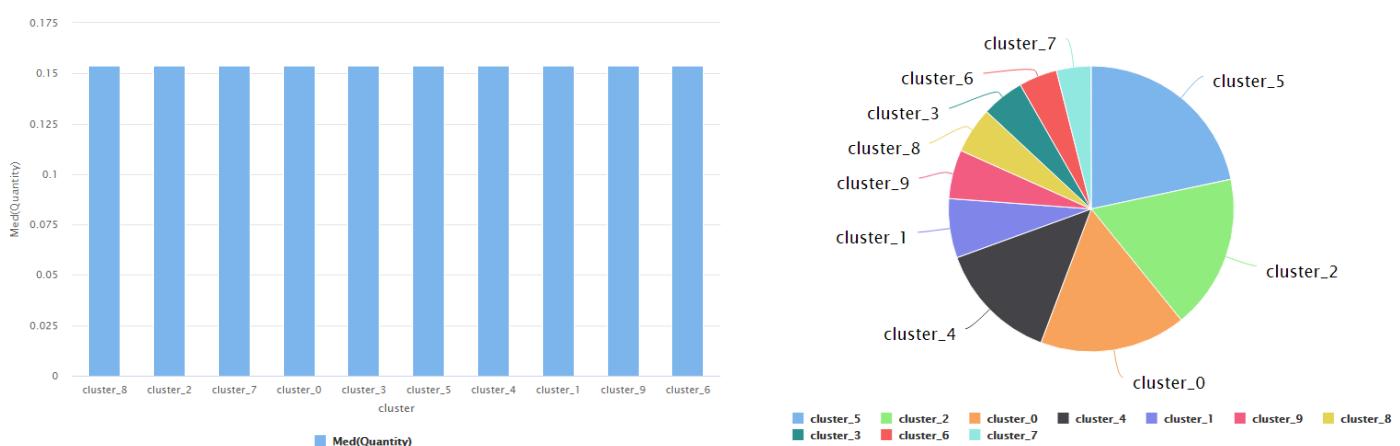
شکل ۶۰ – جدول مراکز خوشه ها



شکل ۶۱ – نقشه حرارتی خوشها

Cluster 0	8,538	Average Distance: 0.141
Order Date is on average 43.74% smaller, New Ship Mode is on average 23.53% larger, New Order Priority is on average 11.89% larger		
Cluster 1	3,478	Average Distance: 0.170
New Ship Mode is on average 88.78% smaller, Order Date is on average 68.78% larger, New Order Priority is on average 33.69% smaller		
Cluster 2	8,774	Average Distance: 0.107
Discount is on average 70.19% smaller, New Ship Mode is on average 21.82% larger, Order Date is on average 12.52% larger		
Cluster 3	2,404	Average Distance: 0.213
New Ship Mode is on average 79.76% smaller, Order Date is on average 43.74% smaller, New Order Priority is on average 40.13% smaller		
Cluster 4	6,868	Average Distance: 0.148
Order Date is on average 100.00% smaller, New Ship Mode is on average 23.27% larger, New Order Priority is on average 11.31% larger		
Cluster 5	11,021	Average Distance: 0.104
Discount is on average 70.17% smaller, Order Date is on average 68.78% larger, New Ship Mode is on average 21.67% larger		
Cluster 6	2,345	Average Distance: 0.119
Discount is on average 259.85% larger, Sales is on average 34.81% smaller, New Ship Mode is on average 20.05% larger		
Cluster 7	2,113	Average Distance: 0.222
Order Date is on average 100.00% smaller, New Ship Mode is on average 78.40% smaller, New Order Priority is on average 44.67% smaller		
Cluster 8	2,665	Average Distance: 0.170
New Ship Mode is on average 88.18% smaller, New Order Priority is on average 33.57% smaller, Order Date is on average 12.52% larger		
Cluster 9	3,006	Average Distance: 0.119
Discount is on average 260.70% larger, Order Date is on average 68.78% larger, Sales is on average 39.02% smaller		

شکل ۶۲ – خلاصه نتایج خوشه ها



شکل ۶۴ – میانه سودآوری خوشه ها

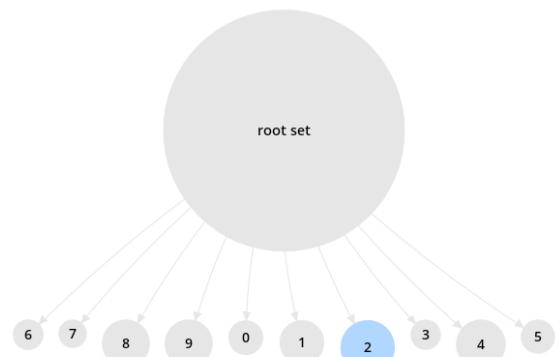
شکل ۶۳ – مجموع سودآوری خوشه ها

همانطور که شکل های بالا نشان می دهند، باز هم بزرگترین خوش بیشترین خوش سودآوری را دارد اما میانه سودآوری خوش ها و در نتیجه مراکز خوش ها از نظر سودآوری نزدیک به هم می باشند که این موضوع نیز نتایج مدلسازی پیشین را نیز تایید می کند. برخلاف مدل پیشین، خوش بزرگ از نظر سودآوری در جایگاه بالای قرار دارد (مدل قبل رتبه ۵ و این مدل رتبه ۲). مبنای مقایسه، مراکز خوش ها می باشند اما باز هم نمی توان از این نکته چشم پوشی کرد که مراکز خوش ها از نظر سودآوری در فاصله نزدیکی قرار دارند.

در نوع دیگر مدلسازی، متغیرهای کیفی اسمی نیز با تخصیص عدد دخیل خواهند شد. این رویکرد رایج نمی باشد اما در صورتی که چنین متغیرهایی وجود داشته باشد و مبنای فاصله، فاصله اقلیدسی باشد خود نرم افزار به صورت خودکار به آن ها عدد تخصیص می دهد. نتایج حاصل به شکل زیر است.

شاخص ارزیابی در این نوع مدل برابر با ۱,۴۶۸ و متوسط فواصل درون خوش ای ۱,۳۱۸ می باشد که از دو مدل پیشین دارای اعتبار کمتری است.

Cluster 0: 4394 items	Cluster 6: 3847 items
Cluster 1: 5681 items	Cluster 7: 3565 items
Cluster 2: 6997 items	Cluster 8: 6116 items
Cluster 3: 3807 items	Cluster 9: 6134 items
Cluster 4: 6361 items	Total number of items: 51212
Cluster 5: 4310 items	



شکل ۶۵ - مقایسه بزرگی مشخصه ها

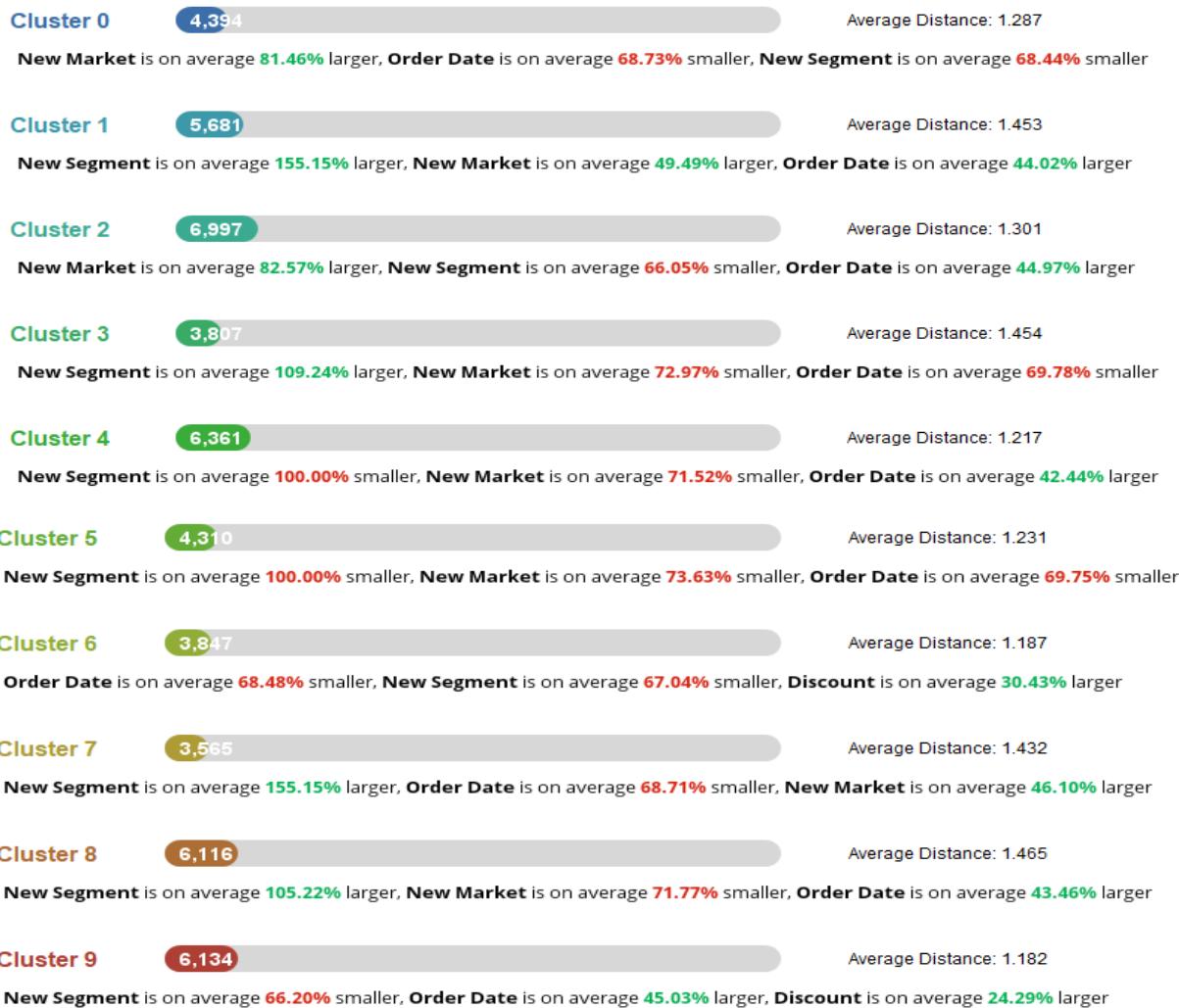
همبستگی میان مشخصه ها مانند دو روش پیشین است.

Attributes	Sales	Quantity	Discount	Profit	Shippin...	Order D...	Ship Date	New Shi...	New Re...	New Order Pri...	New Se...	New Ma...	New Ca...
Sales	1	0.314	-0.087	0.485	0.768	-0.003	-0.003	0.001	-0.004	-0.003	0.002	0.014	0.036
Quantity	0.314	1	-0.020	0.104	0.272	-0.005	-0.005	0.005	-0.002	0.002	-0.001	-0.121	-0.010
Discount	-0.087	-0.020	1	-0.317	-0.078	-0.006	-0.006	-0.010	0.004	0.005	-0.002	-0.042	-0.048
Profit	0.485	0.104	-0.317	1	0.354	0.003	0.002	0.003	-0.001	-0.002	0.001	0.008	0.066
Shipping Cost	0.768	0.272	-0.078	0.354	1	-0.003	-0.004	-0.145	0.003	-0.175	0.002	0.017	0.032
Order Date	-0.003	-0.005	-0.006	0.003	-0.003	1	0.994	-0.002	-0.005	0.014	0.005	0.025	-0.002
Ship Date	-0.003	-0.005	-0.006	0.002	-0.004	0.994	1	0.001	-0.006	0.018	0.006	0.025	-0.002
New Ship Mode	0.001	0.005	-0.010	0.003	-0.145	-0.002	0.001	1	0.007	0.449	0.006	0.004	-0.001
New Returned	-0.004	-0.002	0.004	-0.001	0.003	-0.005	-0.006	0.007	1	0.002	-0.009	-0.013	0.001
New Order Priority	-0.003	0.002	0.005	-0.002	-0.175	0.014	0.018	0.449	0.002	1	0.020	0.005	-0.008
New Segment	0.002	-0.001	-0.002	0.001	0.002	0.005	0.006	0.006	-0.009	0.020	1	0.000	-0.003
New Market	0.014	-0.121	-0.042	0.008	0.017	0.025	0.025	0.004	-0.013	0.005	0.000	1	0.047
New Category	0.036	-0.010	-0.048	0.066	0.032	-0.002	-0.002	-0.001	0.001	-0.008	-0.003	0.047	1

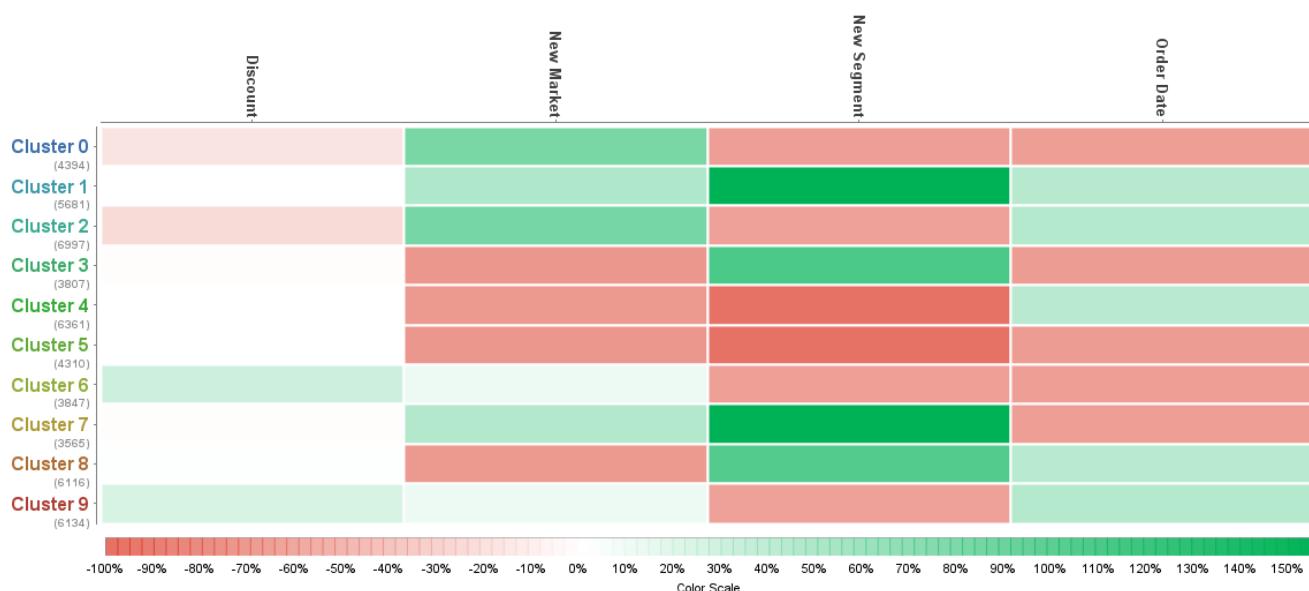
شکل ۶۶ - همبستگی میان مشخصه ها

Cluster	Sales	Quantity	Discount	Profit	Order Date	New Ship M...	New Return...	New Order ...	New Segm...	New Market	New Categ...
Cluster 0	0.011	0.171	0.136	0.442	2012.556	0.753	0.038	0.520	1.247	4.275	2.043
Cluster 1	0.012	0.174	0.169	0.442	2014.560	0.759	0.037	0.538	3	3.698	2.024
Cluster 2	0.011	0.170	0.125	0.442	2014.577	0.750	0.042	0.538	1.266	4.295	2.044
Cluster 3	0.010	0.212	0.166	0.442	2012.537	0.754	0.047	0.533	2.640	1.488	1.968
Cluster 4	0.009	0.208	0.169	0.442	2014.532	0.743	0.041	0.526	1	1.514	1.968
Cluster 5	0.010	0.212	0.168	0.442	2012.538	0.757	0.048	0.527	1	1.476	1.984
Cluster 6	0.012	0.184	0.219	0.442	2012.560	0.750	0.042	0.524	1.258	3	2.015
Cluster 7	0.012	0.176	0.165	0.442	2012.556	0.748	0.042	0.536	3	3.637	2.039
Cluster 8	0.010	0.212	0.170	0.442	2014.550	0.745	0.046	0.533	2.609	1.509	1.972
Cluster 9	0.013	0.188	0.209	0.442	2014.578	0.758	0.049	0.527	1.265	3	2.003

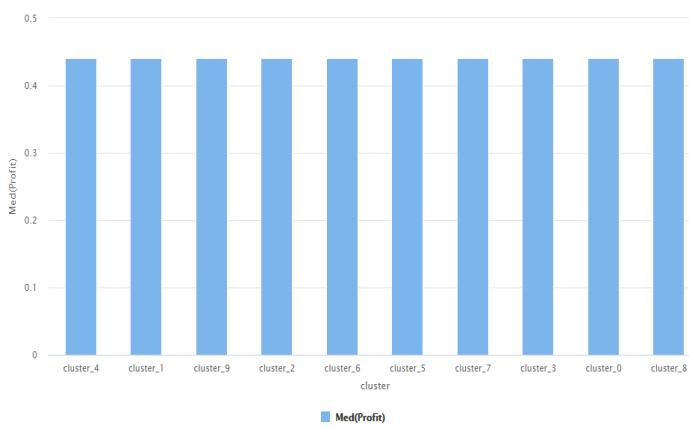
شكل ٦٧ - جدول مراكز خوشة ها



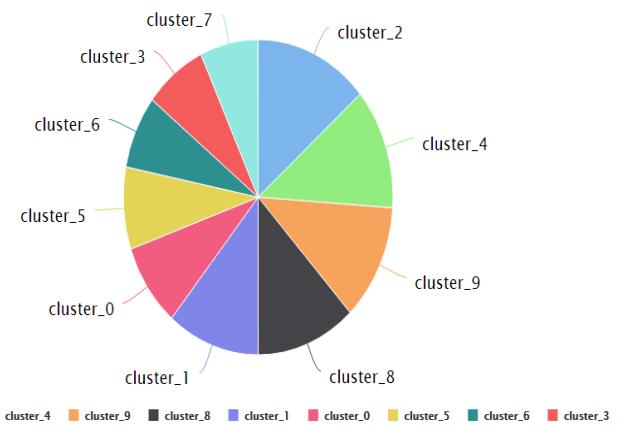
شكل ٦٨ - خلاصه نتایج حاصل از مدل



شکل ۶۹ – نقشه حرارتی خوشه ها



شکل ۷۱ – میانه سودآوری خوشه ها

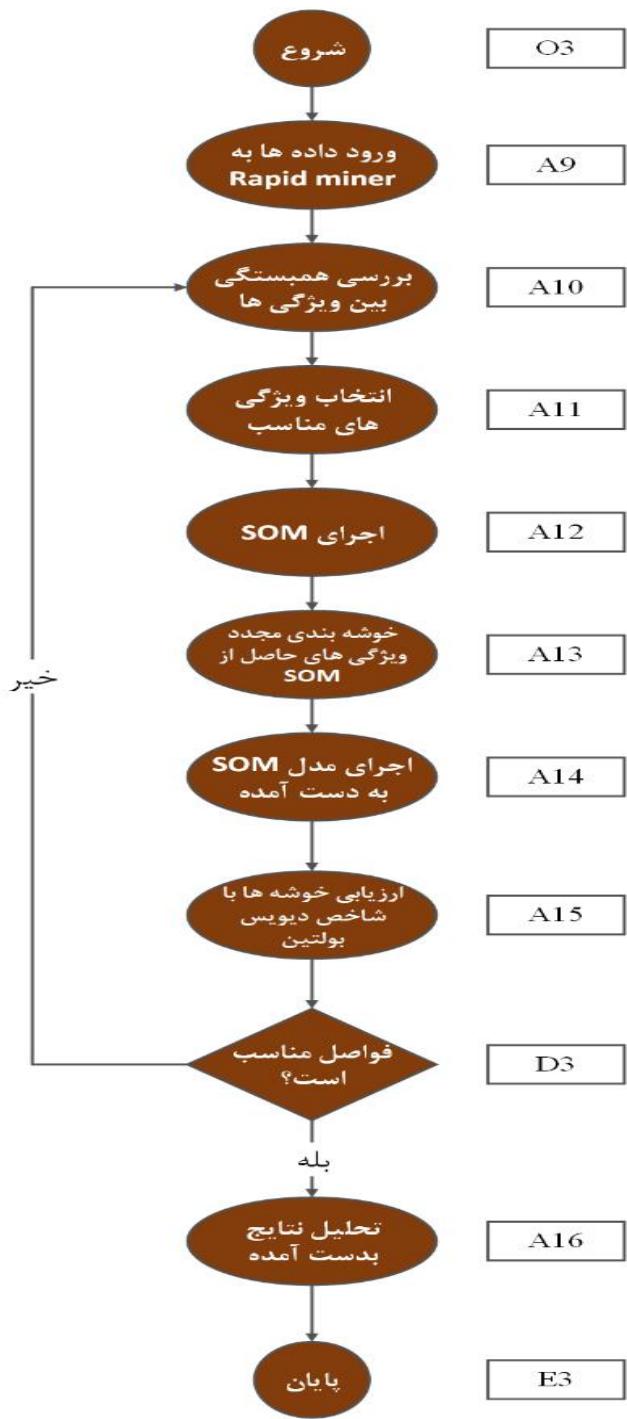


شکل ۷۰ – مجموع سودآوری خوشه ها

در این نوع از مدلسازی نیز تایید شد که مراکز خوشه ها از نظر سودآوری نزدیک می باشند (در این مدل برابر می باشند). تفاوتی که وجود دارد، توزیع متناسب تر رکوردها در هر خوشه می باشد به طوری که از نظر تعداد تفاوت میان خوشه ها از نظر تعداد، متناسب تر از ۲ مدل پیشین می باشد.

همانطور که انتظار می رفت، در حالتی که تنها از مشخصه های کمی استفاده شد به نتایج بهتری دست یافتیم بنابراین نتایج آن مبنای مقایسه در مراحل بعدی قرار خواهند گرفت (هرچند که نزدیکی مراکز خوشه ها از نظر سودآوری در همه حالات، نشان از سیاست مشخصی از شرکت داشت).

گام ۳: نقشه خودسازمانده (SOM)



روش دیگر خوشه بندی، استفاده از نقشه های خودسازمانده می باشد که برای این منظور، در ابتدا برای جلوگیری از تکرار، همبستگی بین ویژگی ها را بررسی کرده و از ویژگی هایی که وزن بالایی در همبستگی دارند صرف نظر خواهد شد. سپس با استفاده از SOM، ویژگی ها استخراج شده و سپس بار دیگر با روش K-میانگین، ویژگی های حاصل خوشه بندی خواهند شد. در نهایت مدل SOM بدست آمده اجرا خواهد شد و در صورت تایید در گام ارزیابی، نتایج حاصل را تحلیل خواهیم کرد.

شکل ۷۲ – فلوچارت نقشه خودسازمانده

در ابتدا داده ها را وارد کرده و مطابق با رویکرد خوش بندی به روش K-میانگین، از عملگر Generate Attribute استفاده کرده و برای متغیرهای اسمی یک عدد تعریف می کنیم و سپس با استفاده از عملگر Select Attribute، مشخصه های مورد نظر را انتخاب می کنیم. نقشه های خودسازمانده روشی برای کاهش بُعد داده ها است و در دسته شبکه های عصبی قرار می گیرد بنابراین می توان از متغیرهای کیفی به عنوان برچسب یا label هر دسته استفاده کرد. در این مدل آزادی عمل بیشتری وجود دارد و می توان هم متغیرهای کیفی را بطور مستقیم و هم با استفاده از نماد عددی به کار برد. مانند رویکرد قبل، این بار نیز روش های مختلف آزمایش شده و بهترین نتیجه انتخاب می شود.

در ادامه، همبستگی بین ویژگی ها بررسی شده و سپس بر مبنای وزن هر ویژگی، مهم ترین آن ها را انتخاب کرده و نمایش می دهیم. همچنین، این ویژگی ها و مقادیر آن ها به عنوان ورودی فرآیند SOM انتخاب شده و نتایج حاصل از SOM علاوه بر مصورسازی، یکبار دیگر با استفاده از روش K-میانگین خوش بندی می شوند. در واقع این خوش بندی بر روی مهم ترین ویژگی ها صورت می گیرد که در مراحل قبل تر انتخاب شده اند. پس از خوش بندی، ارزیابی خوش ها با شاخص دیویس بولدین صورت گرفته و نتایج این ارزیابی نیز نمایش داده می شوند. از عملگر Visualize Model by SOM نیز برای مصور سازی خوش ها استفاده خواهیم کرد.

از آن جا که در بخش مربوط به خوش بندی با K-میانگین(گام ۲) توضیحات مربوط به روش کار و جزئیات هر عملگر به صورت کامل و با شکل نشان داده شد، در این گام از این گام صرف نظر می شود و تنها مدل نهایی نمایش داده خواهد شد و همچنین جزئیات عملگرها و روش مدلسازی نیز بیان می شود. نکاتی که درباره این مدل وجود دارد عبارت است از:

الف) به همان روشی که در خوش بندی برای ویژگی های کیفی یک برچسب عددی تعیین شد، در همین روش نیز این کار انجام می شود(برای داشتن گزینه های مختلف و آزمون و خطأ).

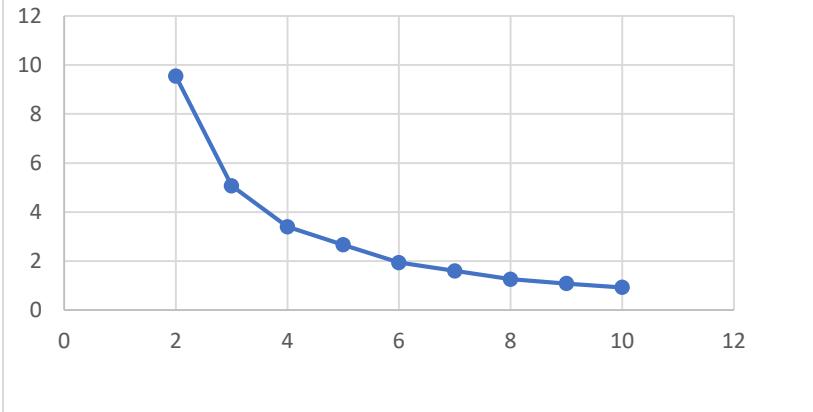
ب) برای جلوگیری از نمایش پیام خطأ توسط نرم افزار، در هر بار لازم است یک متغیر کیفی انتخاب و در گام بعدی به عنوان برچسب یا label انتخاب شود(زیرا نقشه های خودسازمانده روشی مبتنی بر شبکه عصبی می باشند).

ج) ویژگی هایی که وزن آنها ۰,۵ و بیشتر باشد انتخاب خواهند شد.

د) تعداد داده های آموزشی عملگر SOM عدد ۱۰۰ است.

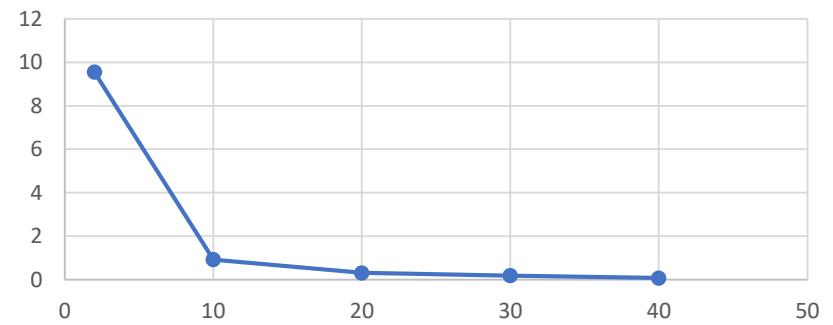
۵) تعداد خوش ها برابر با ۱۰ است(همچنین ۳ نیز گزینه مناسبی است اما ابتدا بهترین حالت یعنی ۱۰ خوش های و سپس ۳ خوش های اجرا خواهد شد).

متوجه فوایل درون خوش ای بر اساس تعداد خوش های مختلف

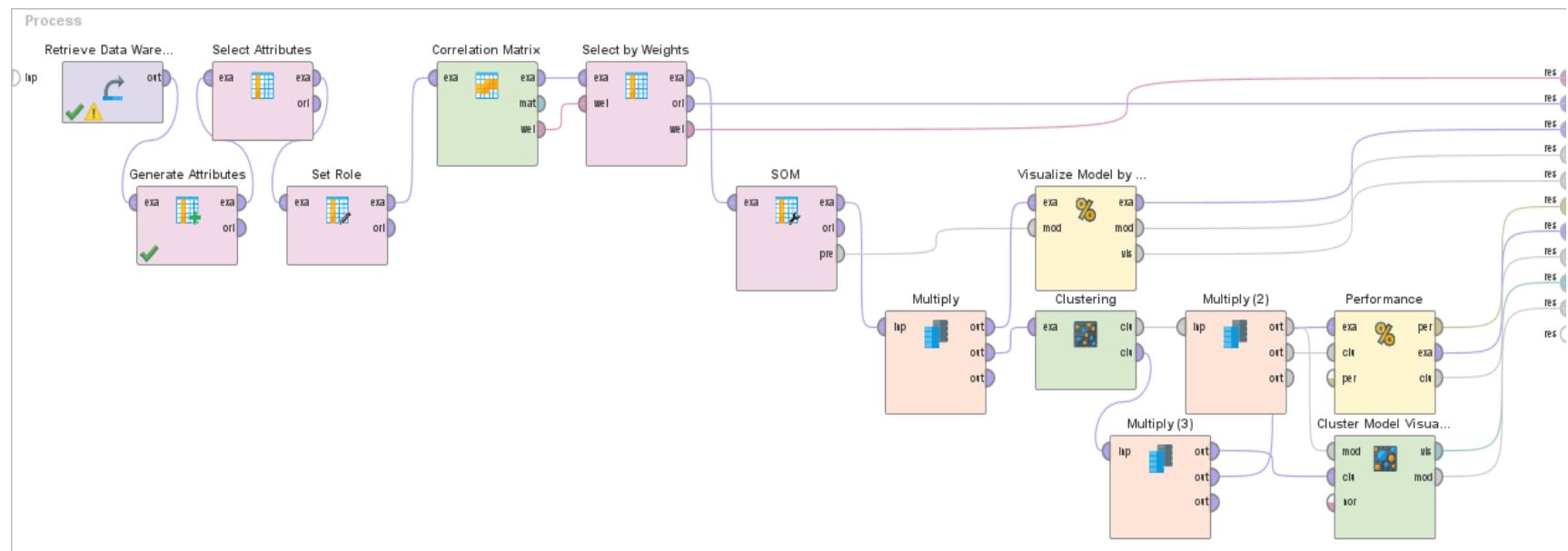


شکل ۷۴ – متوجه فوایل درون خوش ای

متوجه فوایل درون خوش ای بر اساس تعداد خوش های مختلف



شکل ۷۳ – متوجه فوایل درون خوش ای



شکل ۷۵ – مدل SOM

۲-۳- نتایج مدل

در ابتدا، ویژگی بازار به همراه دیگر متغیر های کمی انتخاب خواهند شد. پاسخ به ازای انتخاب هر متغیر کیفی دیگری به عنوان برچسب یکی خواهد شد(از نظر تعداد خوش و شاخص های فاصله) زیرا متغیرهای کمی در همه حالات یکسان می باشند.

میزان شاخص دیویس بولدین برای این مدل عدد ۶۳۰،۰ و متوسط فواصل درون خوش ای ۸۸۳،۰ است که این اعداد نشان از مدلسازی مناسب دارند. علاوه بر این، توزیع متناسب تراکنش ها در هر ۱۰ خوش نیز گواهی بر این موضوع است.

Cluster 0: 6745 items

Cluster 1: 5382 items

Cluster 2: 3863 items

Cluster 3: 4697 items

Cluster 4: 3258 items

Cluster 5: 3325 items

Cluster 6: 5339 items

Cluster 7: 8293 items

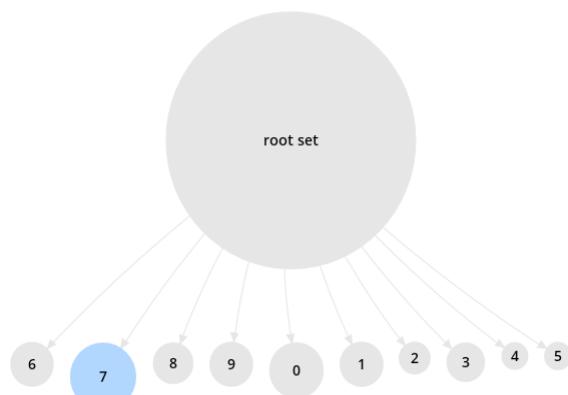
Cluster 8: 4947 items

Cluster 9: 5363 items

Total number of items: 51212

attribute	weight
Sales	0.066
Quantity	0.921
Discount	1
Profit	0.591
Shipping Cost	0.221
Order Date	0.000
Ship Date	0

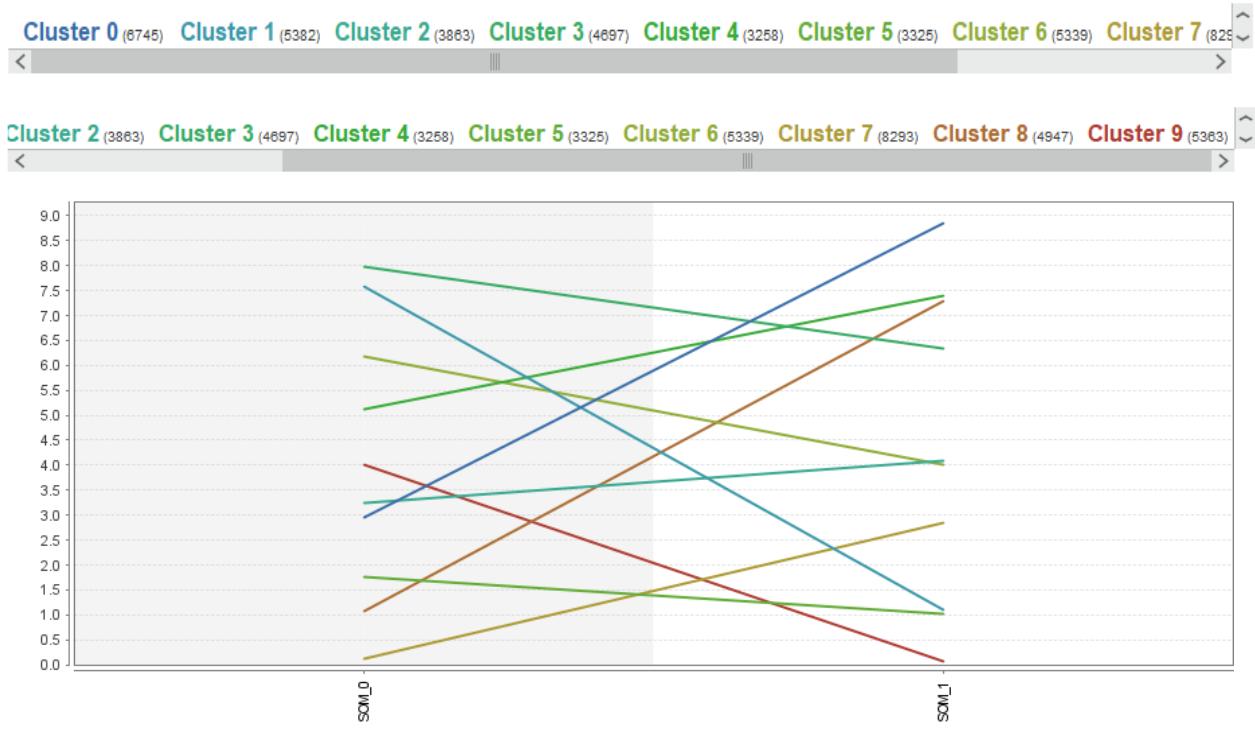
شکل ۷۷ – جدول وزن مشخصه ها



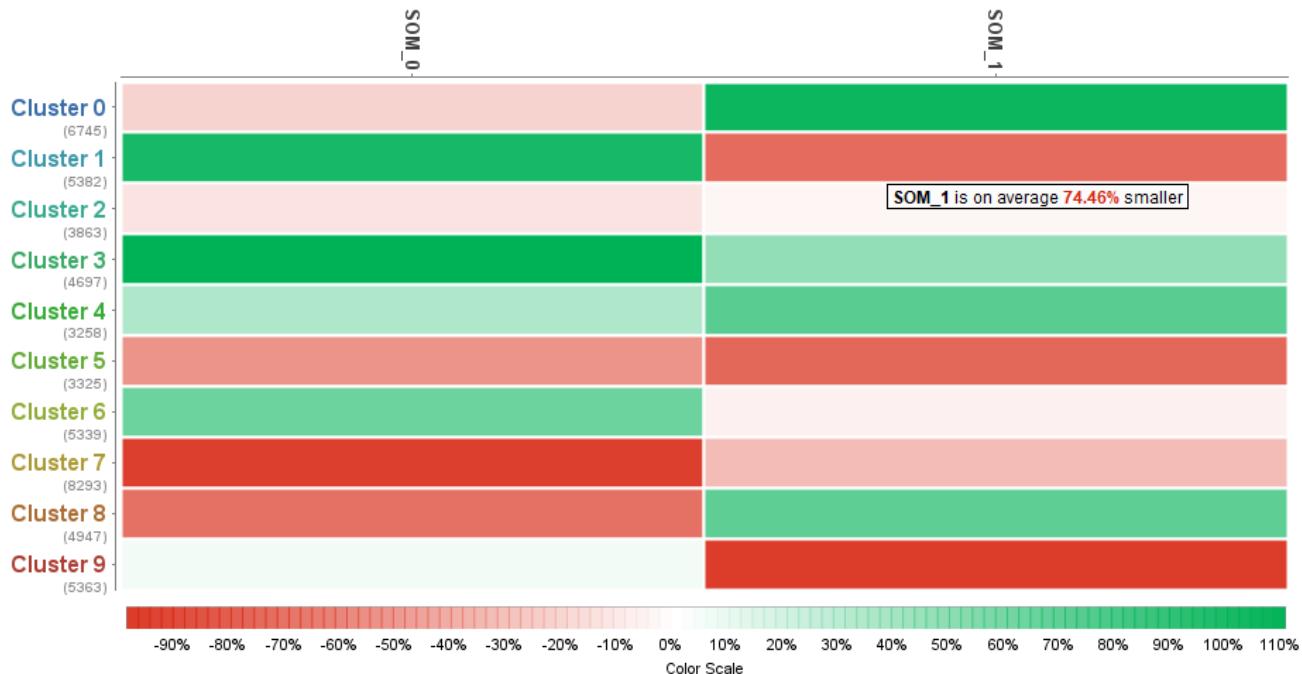
شکل ۷۶ – بزرگی هر خوش

Cluster	SOM_0	SOM_1
Cluster 0	2.940	8.844
Cluster 1	7.584	1.097
Cluster 2	3.245	4.098
Cluster 3	7.986	6.338
Cluster 4	5.110	7.405
Cluster 5	1.748	1.020
Cluster 6	6.188	3.999
Cluster 7	0.123	2.842
Cluster 8	1.071	7.301
Cluster 9	4.001	0.067

شکل ۷۸ – جدول مراکز خوش



شکل ۷۹ - مقایسه مراکز خوشه ها



شکل ۸۰ - نمودار حرارتی خوشه ها

Cluster 0 6,745 Average Distance: 0.352

SOM_1 is on average **105.79%** larger, SOM_0 is on average **22.38%** smaller

Cluster 1 5,382 Average Distance: 0.965

SOM_0 is on average **100.24%** larger, SOM_1 is on average **74.46%** smaller

Cluster 2 3,863 Average Distance: 1.227

SOM_0 is on average **14.31%** smaller, SOM_1 is on average **4.63%** smaller

Cluster 3 4,697 Average Distance: 1.810

SOM_0 is on average **110.84%** larger, SOM_1 is on average **47.48%** larger

Cluster 4 3,258 Average Distance: 1.864

SOM_1 is on average **72.32%** larger, SOM_0 is on average **34.92%** larger

Cluster 5 3,325 Average Distance: 1.064

SOM_1 is on average **76.25%** smaller, SOM_0 is on average **53.85%** smaller

Cluster 6 5,389 Average Distance: 0.646

SOM_0 is on average **63.39%** larger, SOM_1 is on average **6.94%** smaller

Cluster 7 8,293 Average Distance: 0.482

SOM_0 is on average **96.75%** smaller, SOM_1 is on average **33.86%** smaller

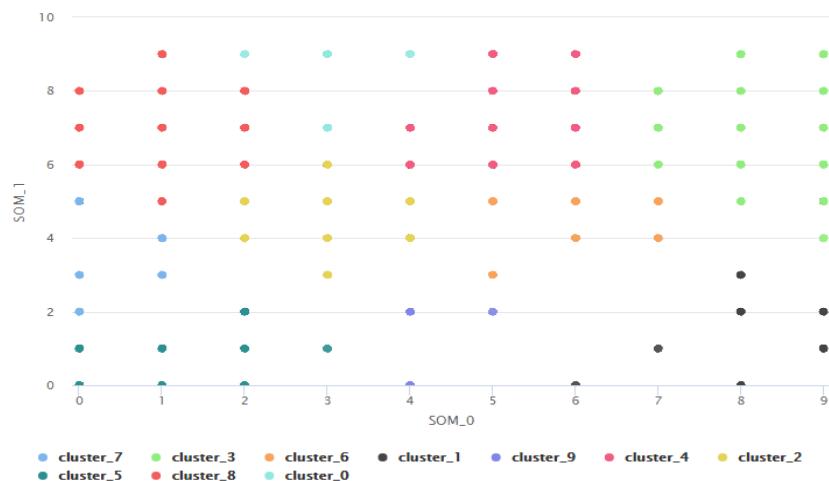
Cluster 8 4,947 Average Distance: 1.349

SOM_0 is on average **71.73%** smaller, SOM_1 is on average **69.89%** larger

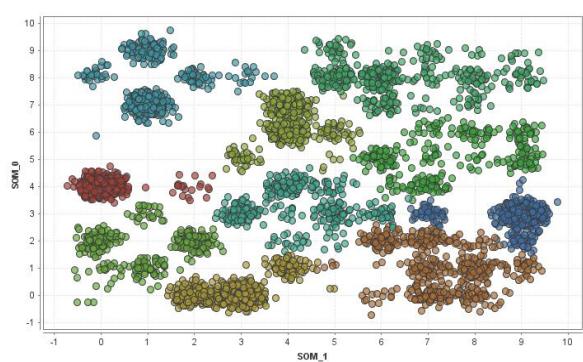
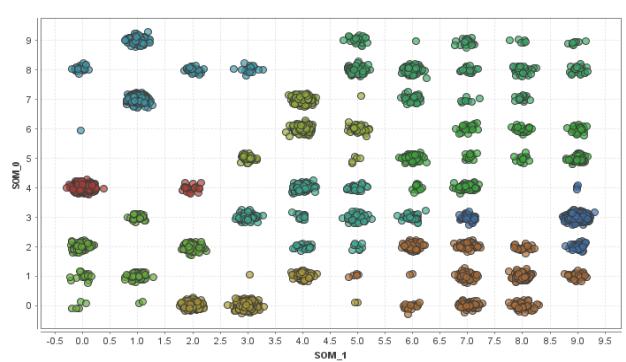
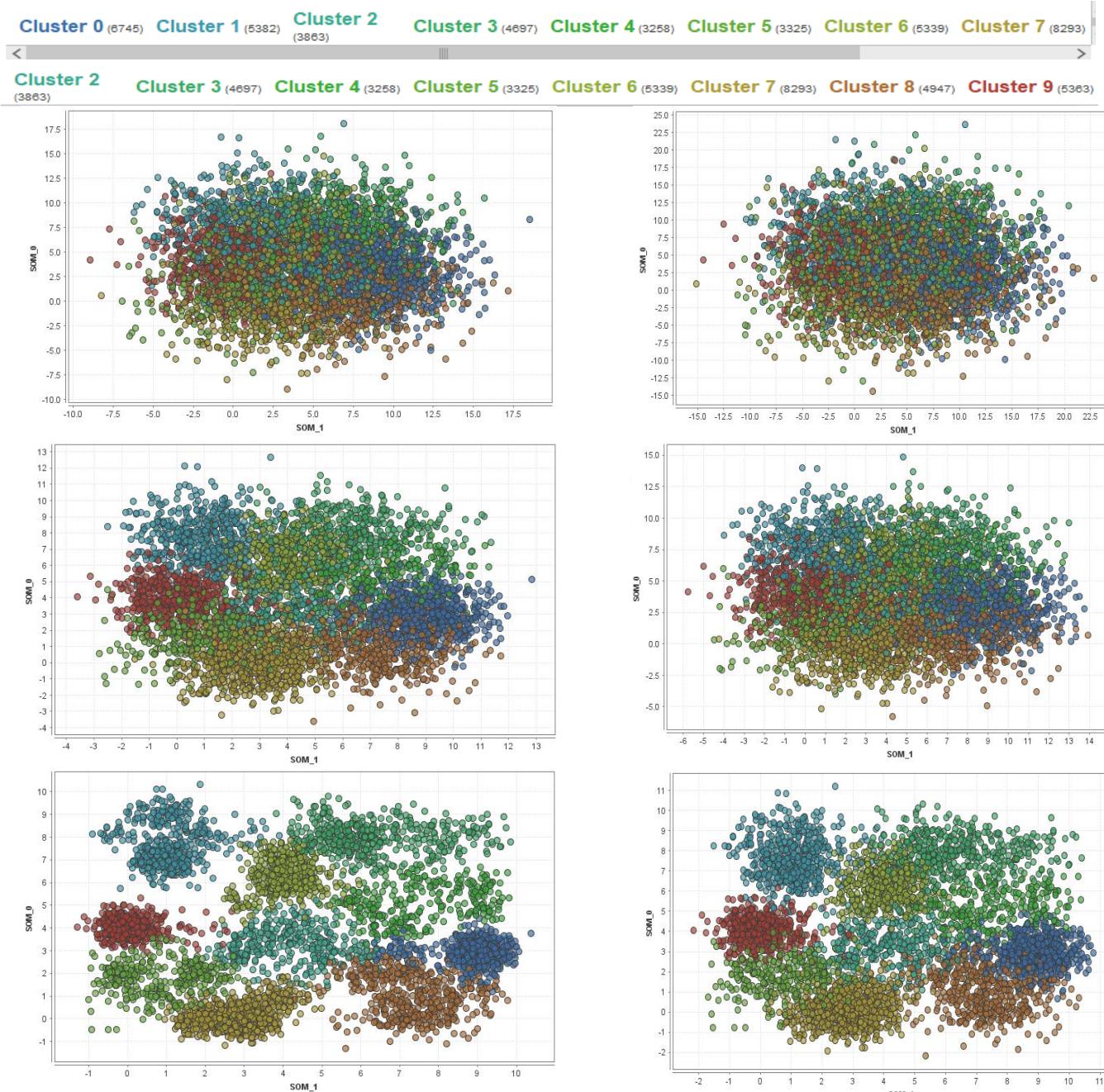
Cluster 9 5,363 Average Distance: 0.131

SOM_1 is on average **98.43%** smaller, SOM_0 is on average **5.63%** larger

شکل ۸۱ – اطلاعات هر خوشه

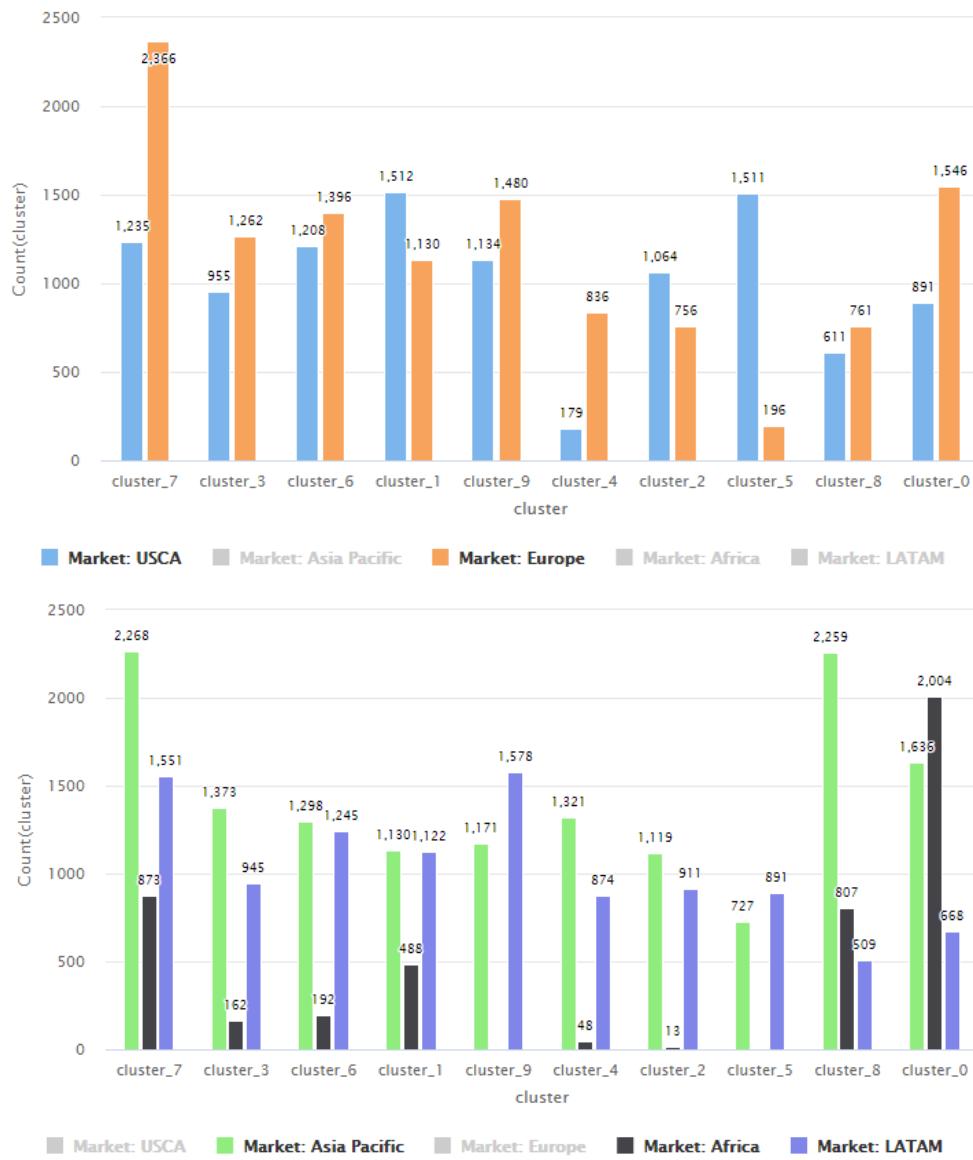


شکل ۸۲ – پراکندگی خوشه ها

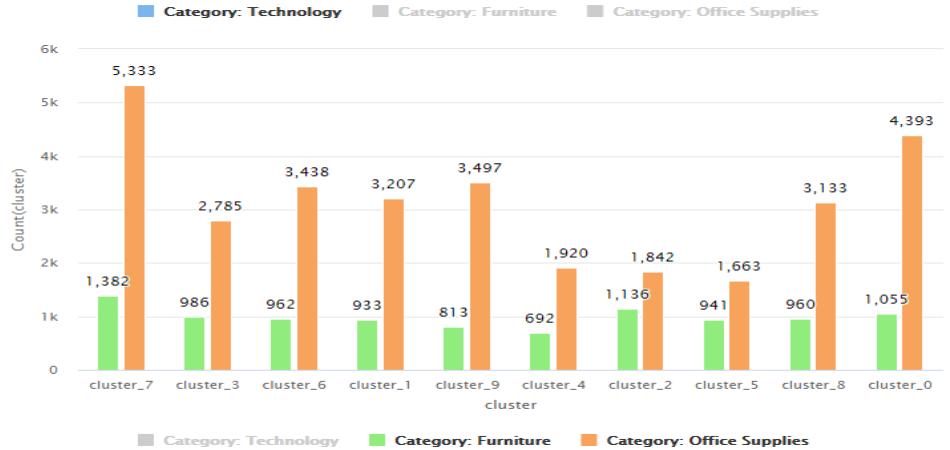
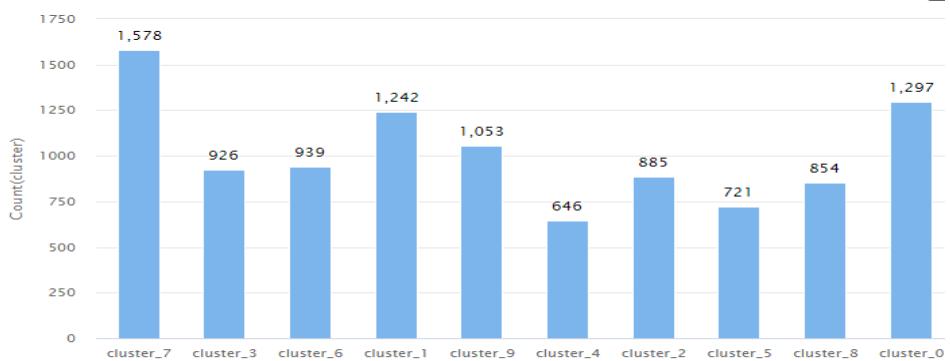


شکل ۸۳ - تخصیص رکوردها به هر خوشه

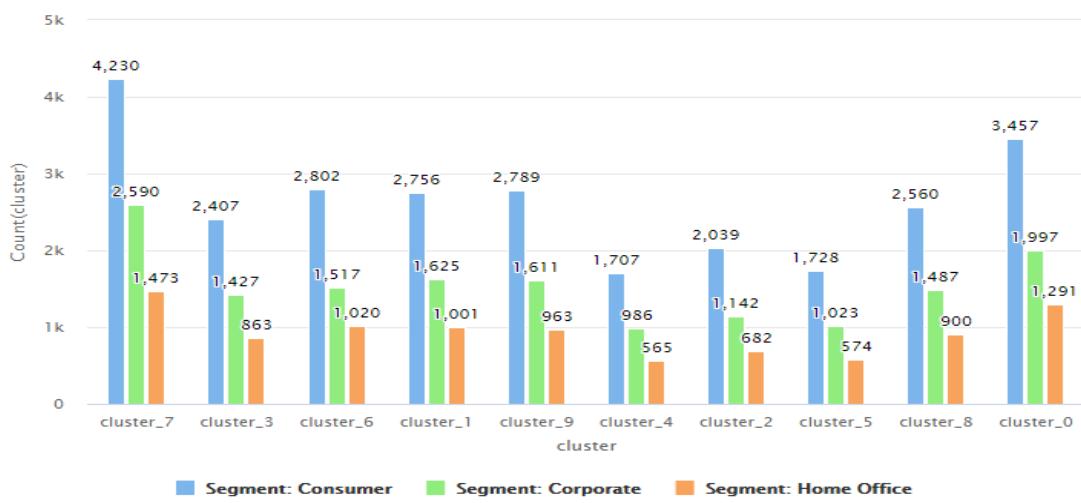
همانطور که در ابتدا نیز گفته شد، خوش بندی با استفاده از برچسب های کیفی دیگر نیز از نظر تعداد اعضای خوش بندی و دیگر مواردی که تاکنون در نمودارها نمایش داده شد یکسان خواهد بود اما مقایسه هر برچسب با خود خوش بندی ها نتایج متفاوتی به دنبال خواهد داشت. برای مثال ممکن است در خوش شماره X بیشترین تراکنش های مربوط به بازار اروپا اما کمترین تراکنش های مربوط به دسته بندی محصولات اداری باشد. بنابراین مدل را یکبار دیگر با استفاده از برچسب های دسته بندی محصول و بخش مشتری (Segment) نیز اجرا کرده و نتایج را مقایسه می کنیم.



شکل ۸۴ – توزیع بازارها در هر خوش



شکل ۸۵ – توزیع دسته بندی محصولات در هر خوشه



شکل ۸۶ – توزیع بخش بندی مشتریان در هر خوشه

در جداول صفحات بعد، خوشه ها بر مبنای مشخصه ها با یکدیگر مقایسه شده و به هر کدام در هر خوشه یک رتبه اختصاص یافته است.

جدول ۳- رتبه بندی خوشه ها بر اساس مشخصه بازار

رتبه در بازار آفریقا	رتبه در بازار آسیا اقیانوسیه	رتبه در بازار اروپا	رتبه در بازار LATAM	رتبه در بازار USCA	رتبه تعداد تراکنش ها	نام خوشه
۱	۳	۲	۹	۸	۲	خوشه ۰
۴	۸	۶	۴	۱	۳	خوشه ۱
۸	۹	۹	۶	۶	۸	خوشه ۲
۶	۴	۵	۵	۷	۷	خوشه ۳
۷	۵	۷	۸	۱۰	۱۰	خوشه ۴
۹	۱۰	۱۰	۷	۲	۹	خوشه ۵
۵	۶	۴	۳	۴	۵	خوشه ۶
۲	۱	۱	۲	۳	۱	خوشه ۷
۳	۲	۸	۱۰	۹	۶	خوشه ۸
۹	۷	۳	۱	۵	۴	خوشه ۹

جدول ۴- رتبه بندی خوشه ها بر اساس مشخصه دسته بندی محصول

رتبه در دسته لوازم اداری	رتبه در دسته بندی مبلمان	رتبه در دسته بندی تکنولوژی	رتبه تعداد تراکنش ها	نام خوشه
۲	۳	۲	۲	خوشه ۰
۵	۸	۳	۳	خوشه ۱
۹	۲	۷	۸	خوشه ۲
۷	۴	۶	۷	خوشه ۳
۸	۱۰	۱۰	۱۰	خوشه ۴
۱۰	۷	۹	۹	خوشه ۵
۴	۵	۵	۵	خوشه ۶
۱	۱	۱	۱	خوشه ۷
۶	۶	۸	۶	خوشه ۸
۳	۹	۴	۴	خوشه ۹

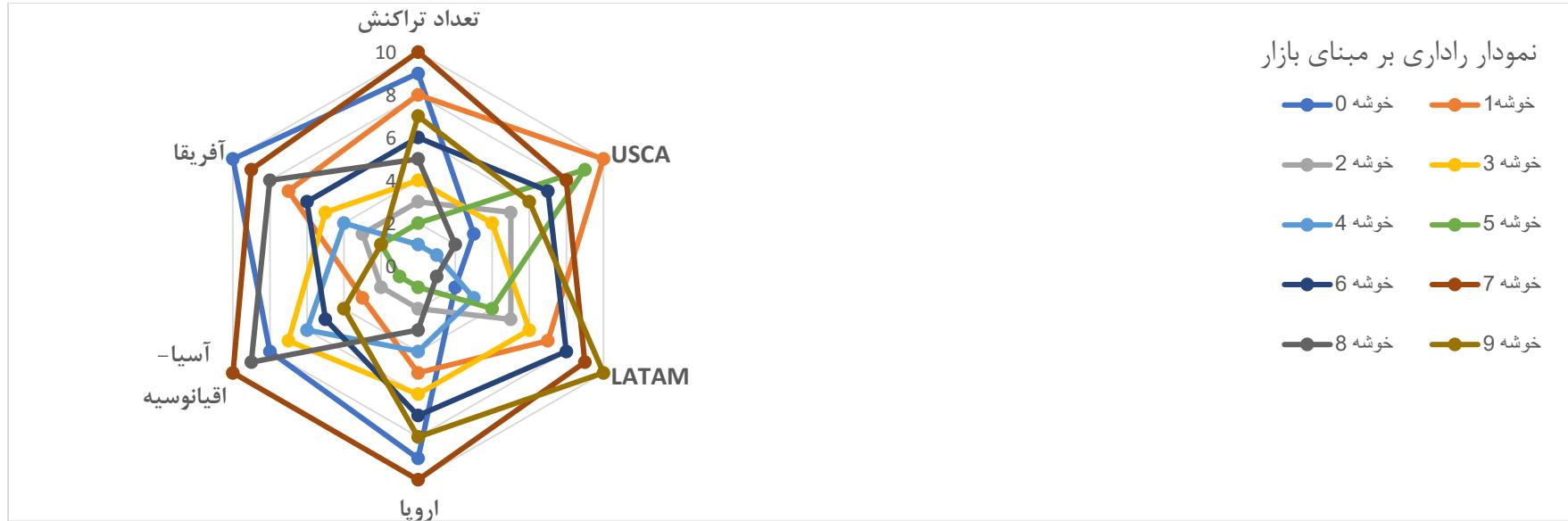
جدول ۵- رتبه بندی خوش‌ها بر اساس مشخصه بخش مشتری

نام خوش	رتبه تعداد تراکنش‌ها	رتبه در بخش مصرف کننده	رتبه در بخش شرکت	رتبه در بخش خانگی
خوشه ۰	۲	۲	۲	۲
خوشه ۱	۳	۵	۳	۴
خوشه ۲	۸	۸	۸	۸
خوشه ۳	۷	۷	۷	۷
خوشه ۴	۱۰	۱۰	۹	۹
خوشه ۵	۹	۹	۵	۱۰
خوشه ۶	۵	۳	۵	۳
خوشه ۷	۱	۱	۱	۱
خوشه ۸	۶	۶	۶	۶
خوشه ۹	۴	۴	۴	۵

اکنون که خوش‌ها به صورت مفصل توضیح داده شدند، مشابه گام ۲ برای هر کدام یک نمودار راداری رسم کرده و سپس مشتریان بر مبنای مشخصات هر کدام از خوش‌ها بخش بندی می‌شوند. نمودارهای صفحات بعد نشان می‌دهد یک خوش معین در زیرمجموعه‌های دسته بندی محصولات و بخش مشتری تفاوت زیادی ندارد. برای مثال اگر در دسته بندی تکنولوژی برتری دارد در دسته بندی لوازم اداری نیز همان برتری را دارد اما در مشخصه بازار قضیه کمی متفاوت است.

جدول ۶- بخش بندی مشتری‌ها

نام خوش/بخش	توضیحات
صفر	بزرگ - دور از بازار آمریکا - رتبه ۲ دسته بندی محصول و بخش مشتری
یک	بزرگ - بازار آمریکا(شمالی و جنوبی) - رتبه ۳(در اغلب موارد) دسته بندی و بخش مشتری
دو	کوچک - رتبه ۲ دسته بندی مبلمان
سه	متوسط - تمایل بیشتری به آسیا و اروپا و آمریکای جنوبی - رتبه ۳ مبلمان
چهار	کوچک - تمایل بیشتر به مبلمان و بخش خانگی
پنج	کوچک - رتبه ۲ بازار آمریکای شمالی - تمایل به مبلمان
شش	متوسط - رتبه ۳ آمریکای لاتین - تمایل به بخش مصرف کننده و خانگی(رتبه ۲ در هر کدام)
هفت	بزرگترین خوش - تمایل به همه بازارها - رتبه ۱ همه دسته بندی‌ها و بخش‌های مشتری
هشت	متوسط - تمایل به آسیا و آفریقا
نه	بزرگ - رتبه ۱ آمریکای جنوبی - تمایل کم نسبت به مبلمان

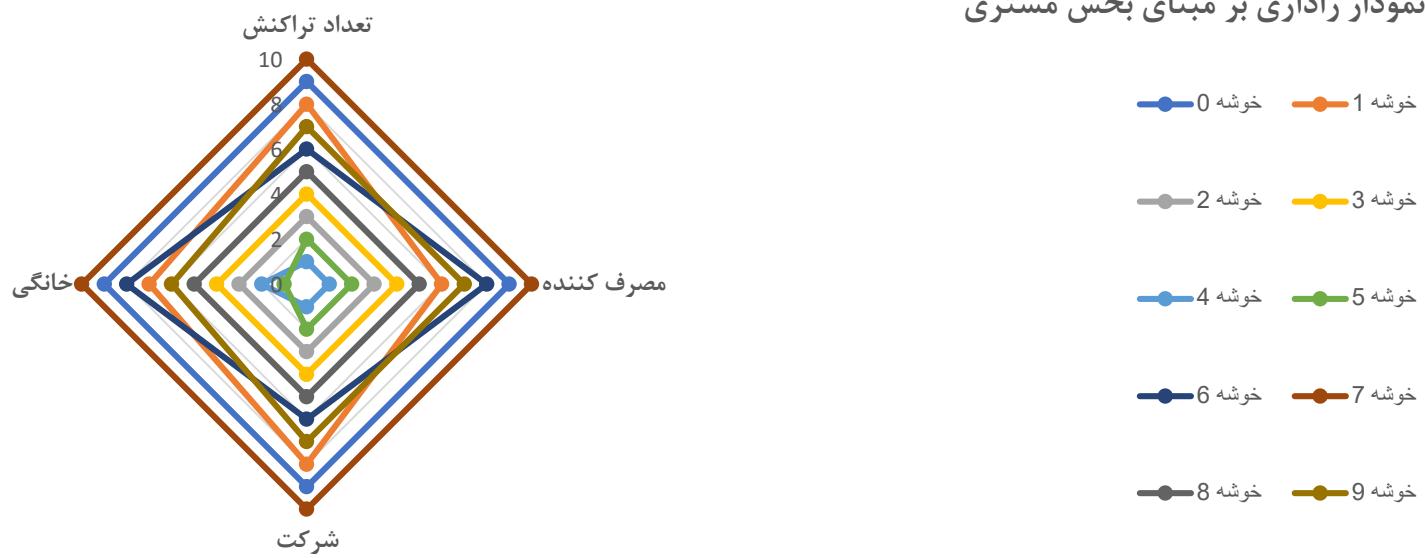


شکل ۸۷ – مقایسه خوش‌های در هر بازار



شکل ۸۸ – مقایسه خوش‌های در هر دسته بندی

نمودار راداری بر مبنای بخش مشتری



شکل ۸۹ – مقایسه خوشه ها در هر بخش مشتری

۳-۳-۳- مدل های دیگر

۱-۳-۳- مدل به همراه متغیرهای کیفی

اکنون، متغیرهای کیفی رتبه ای و متغیر صفر و یک بازگشت (Returned) نیز وارد مدل می شوند.

در این نوع مدلسازی، شاخص دیویس بولدین ۷۲۴ و متوسط فواصل درون خوشه ای ۱,۲۱۷ است بنابراین پاسخ ها معتبر می باشد.

Cluster 0: 6029 items

Cluster 1: 6077 items

Cluster 2: 2944 items

Cluster 3: 6744 items

Cluster 4: 5323 items

Cluster 5: 4874 items

Cluster 6: 4831 items

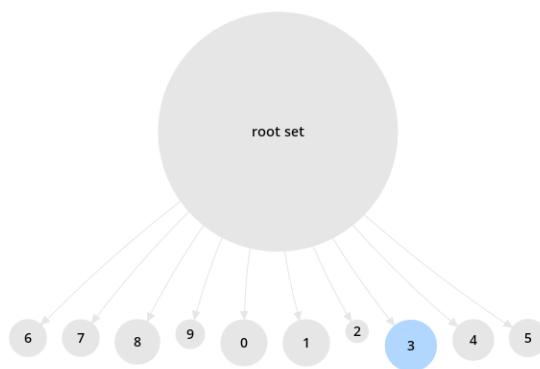
Cluster 7: 4796 items

Cluster 8: 5848 items

Cluster 9: 3746 items

Total number of items: 51212

attribute	weight
Sales	0.059
Quantity	0.814
Discount	0.884
Profit	0.523
Shipping Cost	0.143
Order Date	0.000
Ship Date	0
New Returned	1
New Ship Mode	0.774
New Order Priority	0.764



شکل ۹۰ - بزرگی خوشه ها

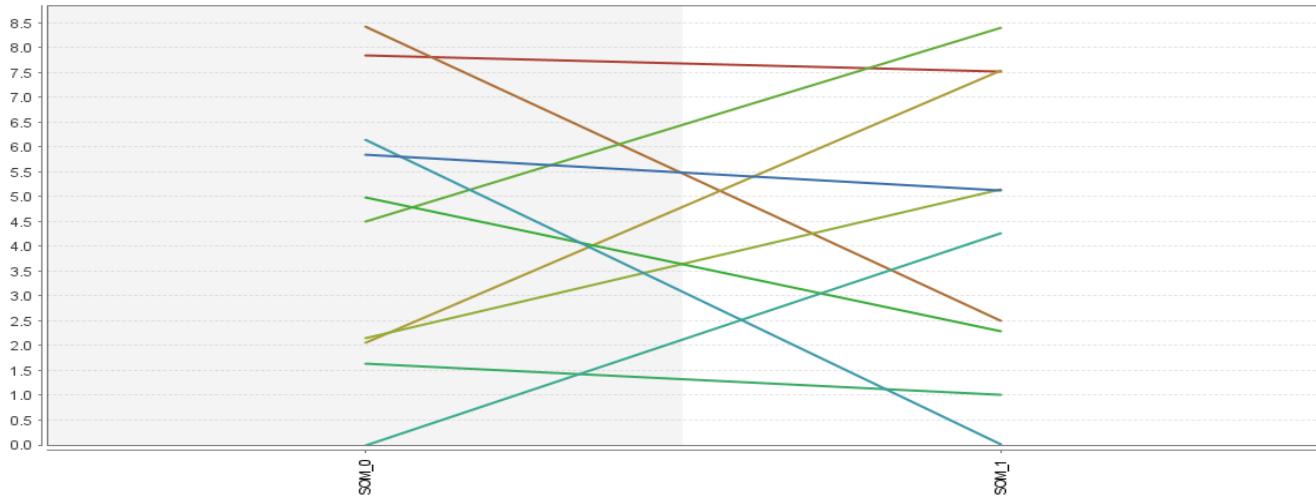
شکل ۹۱ - جدول وزن مشخصه ها

Cluster	SOM_0	SOM_1
Cluster 0	5.852	5.132
Cluster 1	6.160	0.003
Cluster 2	0	4.261
Cluster 3	1.636	1.018
Cluster 4	4.988	2.280
Cluster 5	4.505	8.395
Cluster 6	2.141	5.159
Cluster 7	2.062	7.539
Cluster 8	8.436	2.495
Cluster 9	7.844	7.519

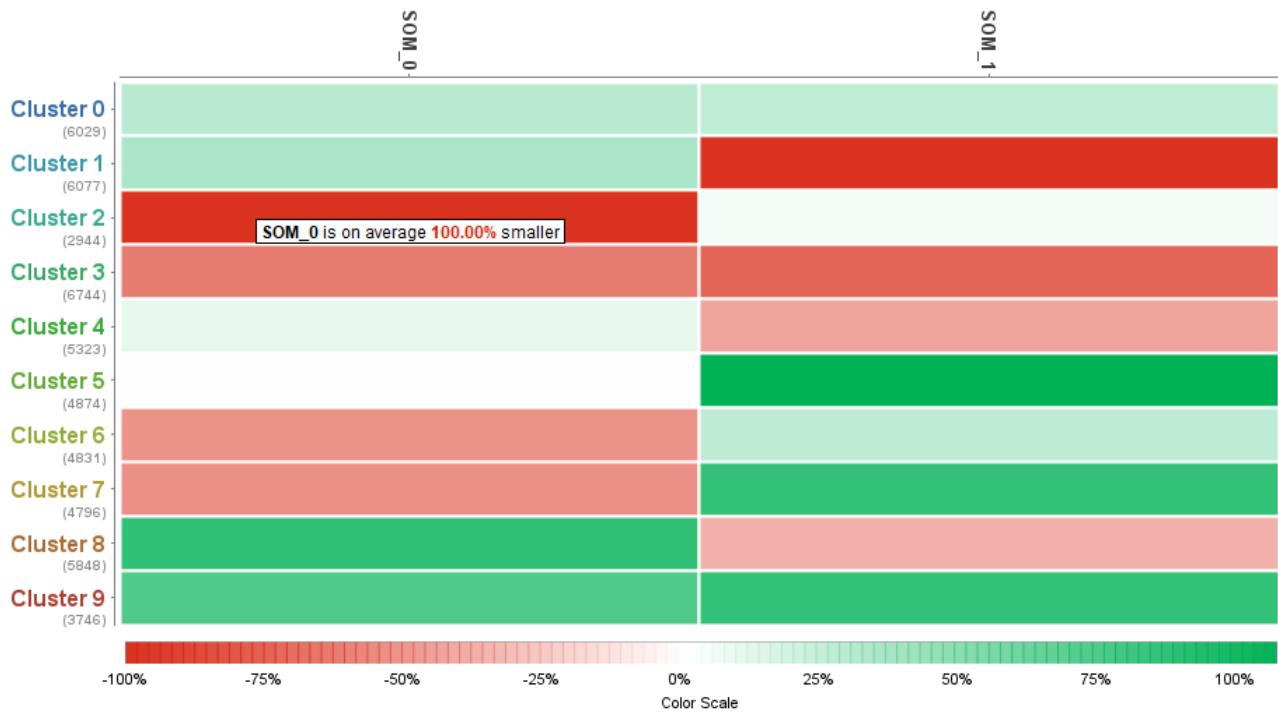
شکل ۹۲ - جدول مراکز خوشه ها

Cluster 0 (6029) Cluster 1 (6077) Cluster 2 (2944) Cluster 3 (6744) Cluster 4 (5323) Cluster 5 (4874) Cluster 6 (4831) Cluster 7 (479)

Cluster 2 (2944) Cluster 3 (6744) Cluster 4 (5323) Cluster 5 (4874) Cluster 6 (4831) Cluster 7 (4796) Cluster 8 (5848) Cluster 9 (3746)



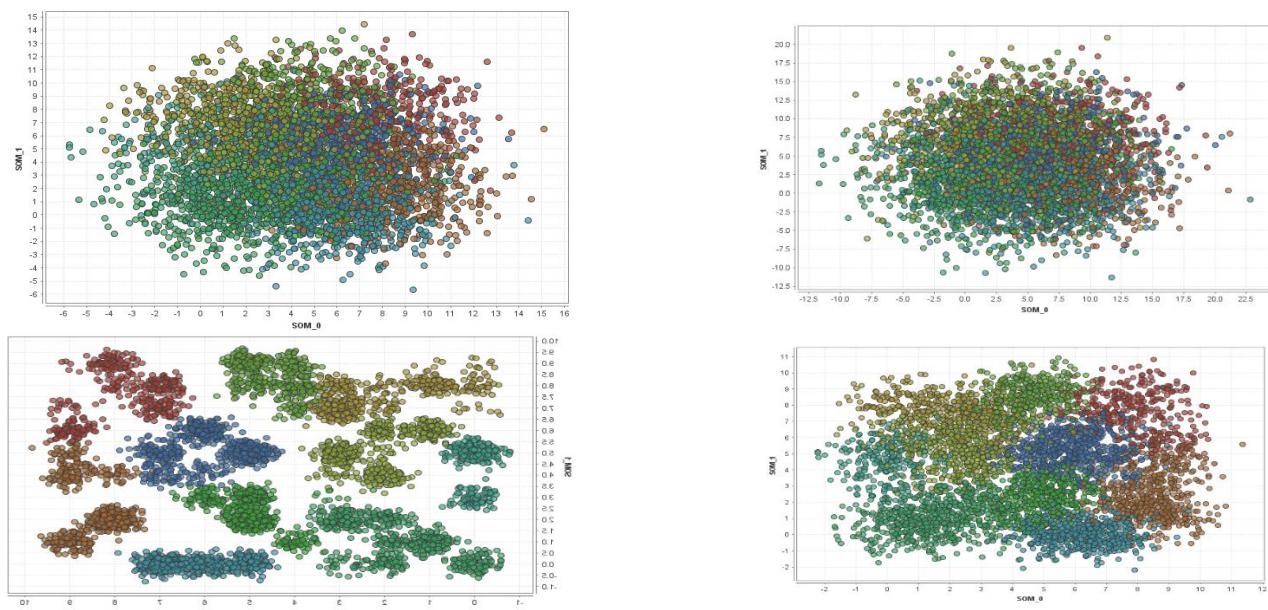
شکل ۹۳ - مرکز خوشه ها



شکل ۹۴ - نقشه حرارتی خوشه ها

Cluster 0	6,029	Average Distance: 1.031
SOM_0 is on average 29.63% larger, SOM_1 is on average 26.81% larger		
Cluster 1	6,077	Average Distance: 0.624
SOM_1 is on average 99.93% smaller, SOM_0 is on average 36.44% larger		
Cluster 2	2,944	Average Distance: 0.932
SOM_0 is on average 100.00% smaller, SOM_1 is on average 5.28% larger		
Cluster 3	6,744	Average Distance: 1.710
SOM_1 is on average 74.85% smaller, SOM_0 is on average 63.77% smaller		
Cluster 4	5,323	Average Distance: 0.833
SOM_1 is on average 43.67% smaller, SOM_0 is on average 10.49% larger		
Cluster 5	4,874	Average Distance: 0.681
SOM_1 is on average 107.44% larger, SOM_0 is on average 0.22% smaller		
Cluster 6	4,831	Average Distance: 1.203
SOM_0 is on average 52.58% smaller, SOM_1 is on average 27.48% larger		
Cluster 7	4,796	Average Distance: 1.534
SOM_1 is on average 86.29% larger, SOM_0 is on average 54.33% smaller		
Cluster 8	5,848	Average Distance: 1.779
SOM_0 is on average 86.86% larger, SOM_1 is on average 38.36% smaller		
Cluster 9	3,746	Average Distance: 1.791
SOM_1 is on average 85.80% larger, SOM_0 is on average 73.75% larger		

شکل ۹۵ – خلاصه اطلاعات خوشه ها



شکل ۹۶ – تخصیص رکوردها به هر خوشه

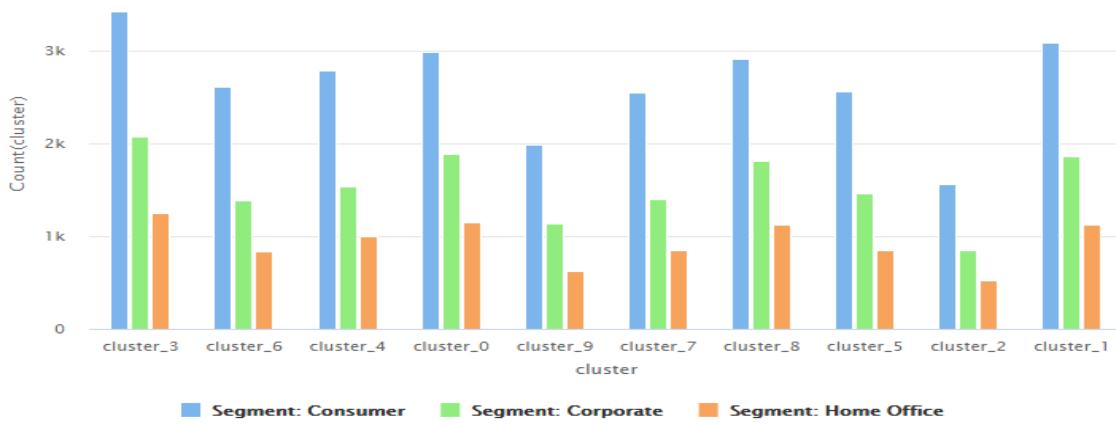
اکنون مطابق روش قبل، هر کدام از متغیرهای کیفی بازار، دسته بندی محصول و بخش بندی مشتری را در هر خوشه نمایش می‌دهیم.



شکل ۹۷ – توزیع بازارها در هر خوشه



شکل ۹۸ – توزیع دسته بندی محصولات در هر خوشه



شکل ۹۹ – توزیع بخش‌ها در هر خوشه

با تفاوتی جزئی، نمودارها شباهت زیادی به نمودارهای پیشین دارند و تنها تفاوت در نام گذاری خوشه ها است. برای مثال خوشه ای که در این مدل در بازار آسیا-اقیانوسیه و آمریکای جنوبی پیشتاز است در بازار آمریکای جنوبی نیز تقریبا همین جایگاه را دارد(در مدل پیشین با اختلاف بسیار کم دوم است). در اشکالی که خلاصه خوشه ها را نمایش می دهند نیز این مورد مشهود است.

۲-۳-۳- مدل با سه خوشه

همانطور که در ابتدا مشاهده شده، به ازای $K=3$ نیز در نمودار شکست (آرنج = Elbow) رخ می دهد بنابراین در آخرین مرحله از این گام، خوشه بندی را با ۳ خوشه نیز انجام داده و به نوعی خوشه ها/بخش های مشتری را در هم ادغام می کنیم. نتایج به صورت زیر است.

شاخص ارزیابی برابر با ۰,۷۹۱ و متوسط فواصل درون خوشه ای ۵,۵۴۱ است.

Cluster 0: 19215 items

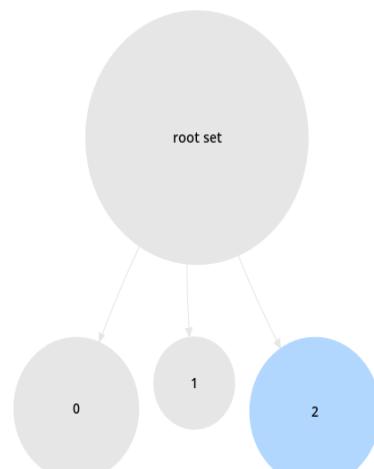
Cluster 1: 11779 items

Cluster 2: 20218 items

Total number of items: 51212

attribute	weight
Sales	0.059
Quantity	0.814
Discount	0.884
Profit	0.523
Shipping Cost	0.143
Order Date	0.000
Ship Date	0
New Returned	1
New Ship Mode	0.774
New Order Priority	0.764

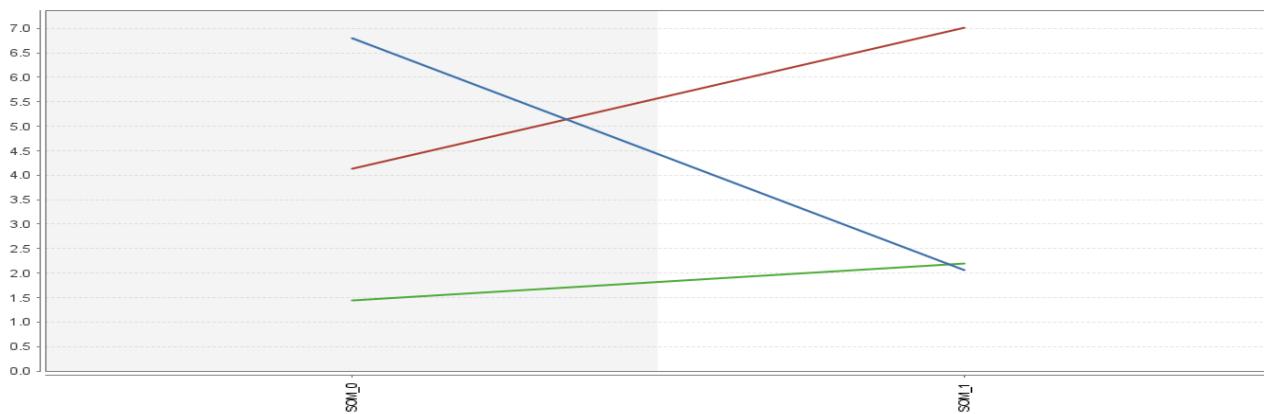
شكل ۱۰۱ – جدول وزن خوشه ها



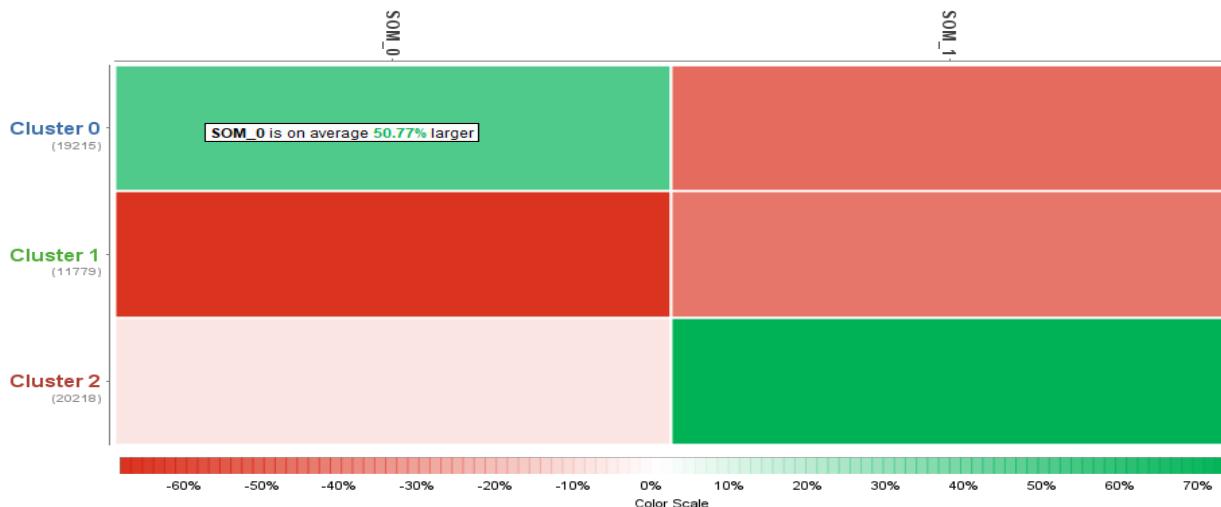
شكل ۱۰۰ – بزرگی خوشه ها

Cluster	SOM_0	SOM_1
Cluster 0	6.807	2.057
Cluster 1	1.435	2.188
Cluster 2	4.130	7.022

شكل ۱۰۲ – مراکز خوشه ها



شکل ۱۰۳ - مراکز خوشه ها



شکل ۱۰۴ - نمودار حرارتی خوشه ها

Cluster 0 19,215 Average Distance: 5.363

SOM_0 is on average **50.77%** larger, **SOM_1** is on average **49.18%** smaller

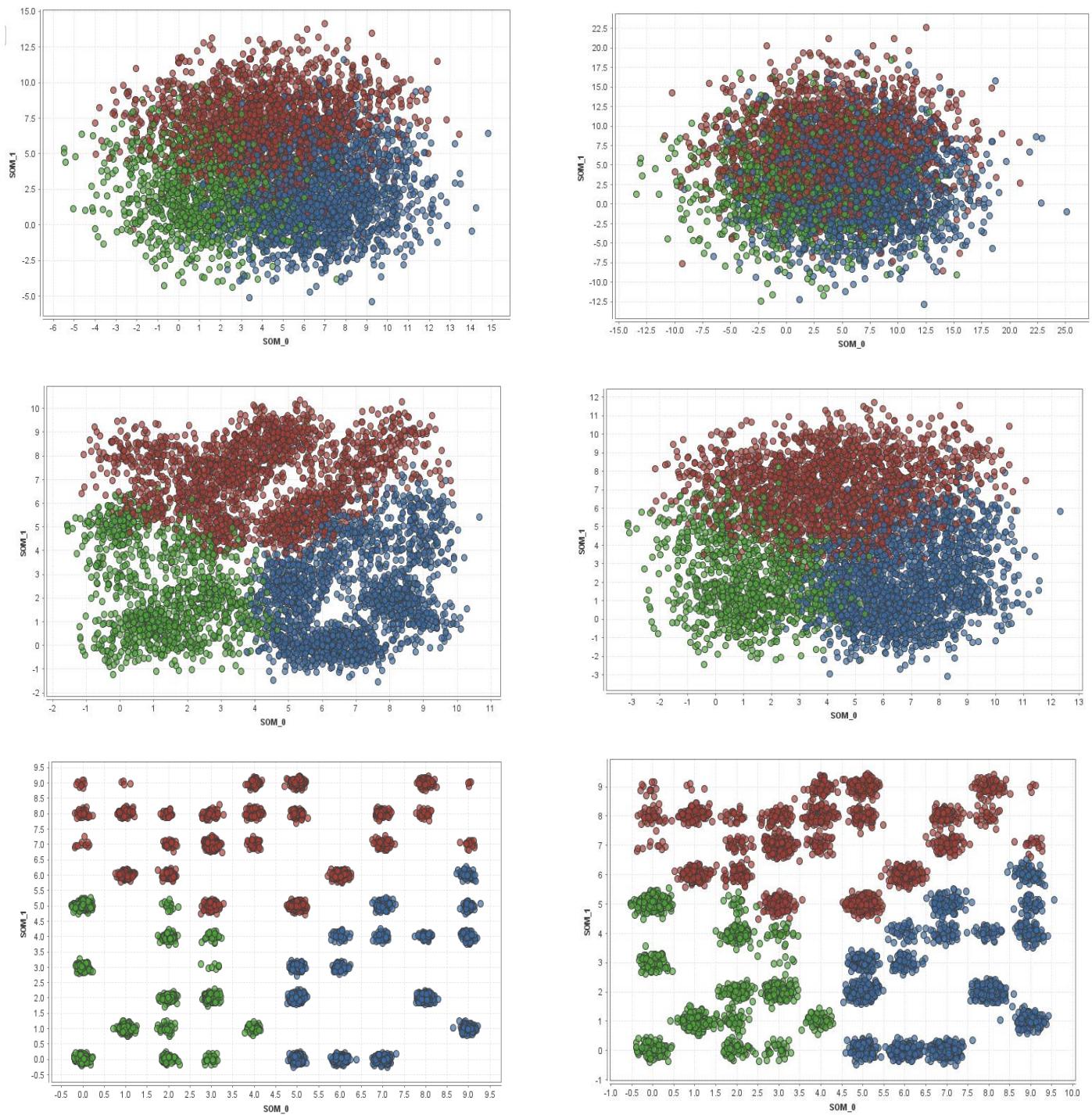
Cluster 1 11,779 Average Distance: 4.608

SOM_0 is on average **68.21%** smaller, **SOM_1** is on average **45.94%** smaller

Cluster 2 20,218 Average Distance: 6.253

SOM_1 is on average **73.51%** larger, **SOM_0** is on average **8.51%** smaller

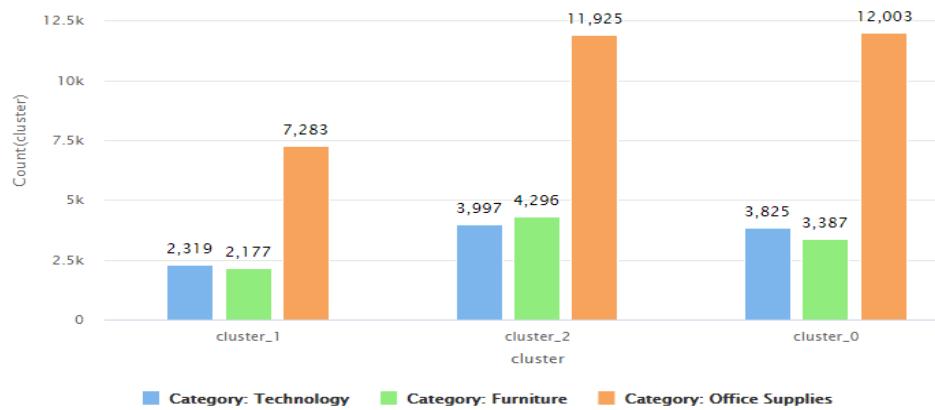
شکل ۱۰۵ - خلاصه اطلاعات خوشه ها



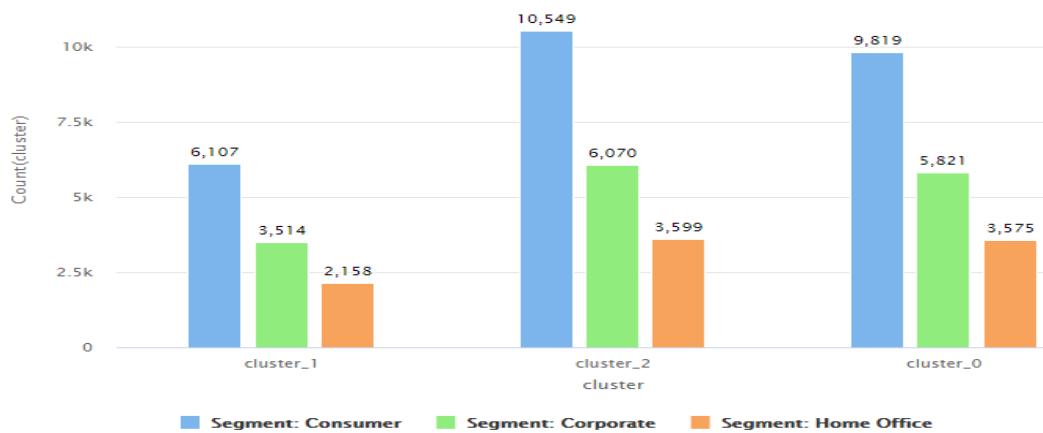
شکل ۱۰۶ – تخصیص رکوردها به هر خوش



شکل ۱۰۷ - توزیع بازارها در هر خوشه



شکل ۱۰۸ - توزیع دسته بندی های محصول در هر خوشه



شکل ۱۰۹ - توزیع بخش ها در هر خوشه

اکنون جدول تخصیص رتبه را مجدد تشکیل می‌دهیم.

جدول ۷- رتبه بندی خوشه ها بر اساس مشخصه بازار

رتبه در بازار آفریقا	رتبه در بازار آسیا اقیانوسیه	رتبه در بازار اروپا	رتبه در بازار LATAM	رتبه در بازار USCA	رتبه تعداد تراکنش ها	نام خوشه
۱	۲	۱	۲	۲	۲	خوشه ۰
۳	۳	۳	۳	۳	۳	خوشه ۱
۲	۱	۲	۱	۱	۱	خوشه ۲

جدول ۸- رتبه بندی خوشه ها بر اساس مشخصه دسته بندی محصول

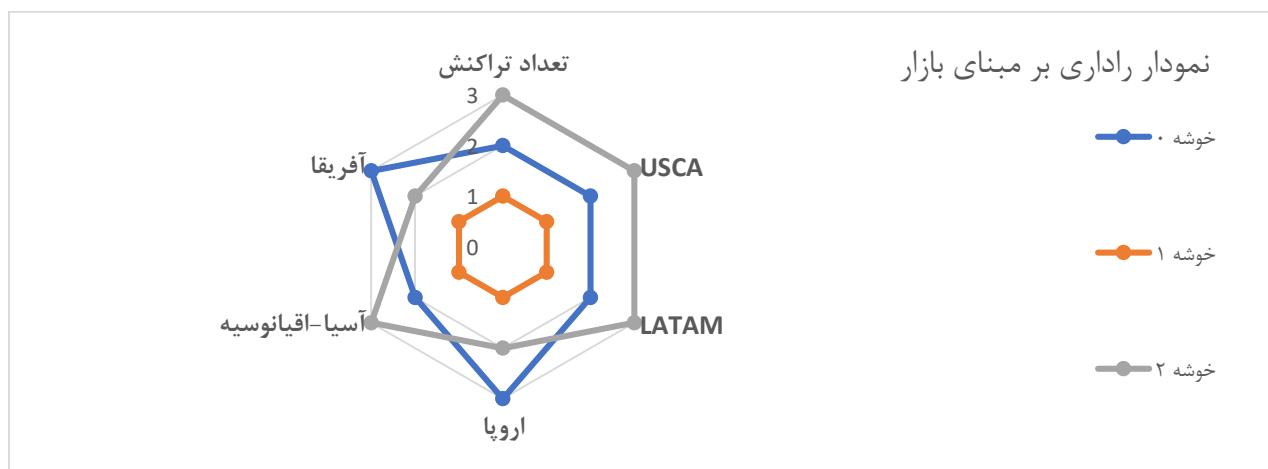
رتبه در دسته لوازم اداری	رتبه در دسته بندی مبلمان	رتبه در دسته بندی تکنولوژی	رتبه تعداد تراکنش ها	نام خوشه
۱	۲	۲	۲	خوشه ۰
۳	۳	۳	۳	خوشه ۱
۲	۱	۱	۱	خوشه ۲

جدول ۹- رتبه بندی خوشه ها بر اساس مشخصه بخش مشتری

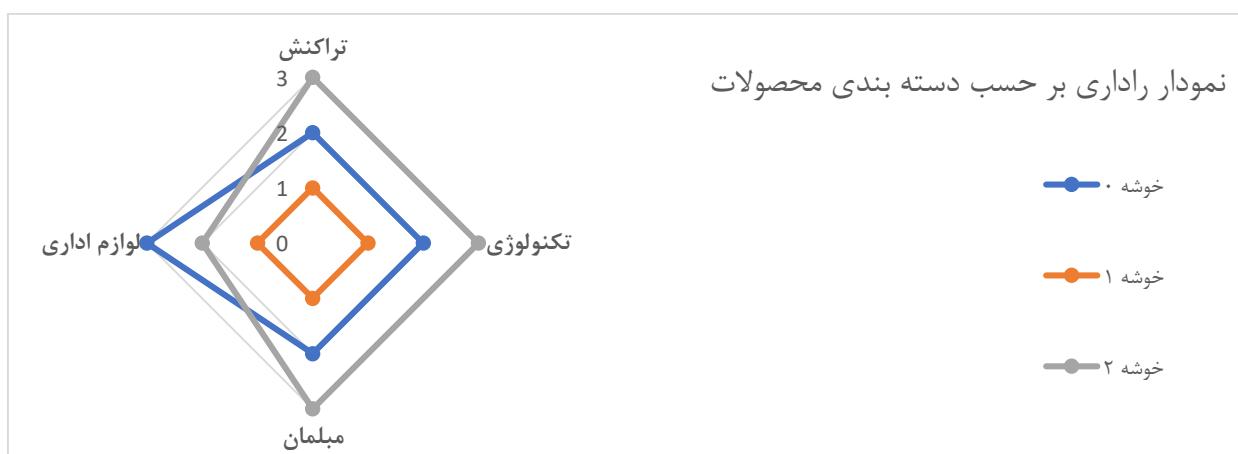
رتبه در بخش خانگی	رتبه در بخش شرکت	رتبه در بخش مصرف کننده	رتبه تعداد تراکنش ها	نام خوشه
۲	۲	۲	۲	خوشه ۰
۳	۳	۳	۳	خوشه ۱
۱	۱	۱	۱	خوشه ۲

جدول ۱۰- بخش بندی مشتری ها

نام خوشه/بخش	توضیحات
صفر	دومین خوشه بزرگ - بازار رتبه ۱ آفریقا و اروپا — رتبه ۱ لوازم اداری — رتبه ۲ همه بخش های مشتری
یک	سومین خوشه بزرگ - در هر بازار و دسته بندی و بخش کوچکترین - بدون تمایلی خاص
دو	بزرگترین خوشه - بازار شماره ۱ آسیا و آمریکا - تمایل کمتر به لوازم خانگی - رتبه ۱ در همه بخش های مشتری

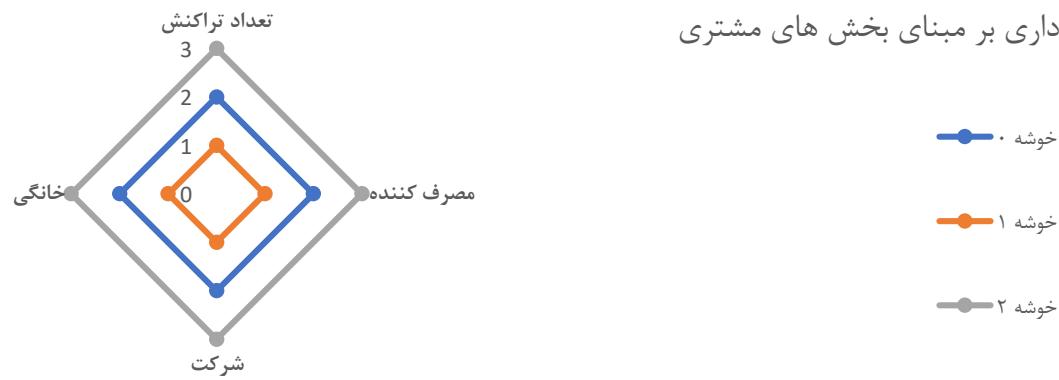


شکل ۱۱۰ - مقایسه خوشه ها در هر بازار



شکل ۱۱۱ - مقایسه خوشه ها در هر دسته محصول

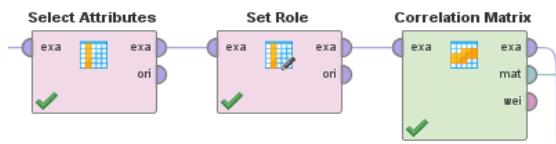
نمودار راداری بر مبنای بخش های مشتری



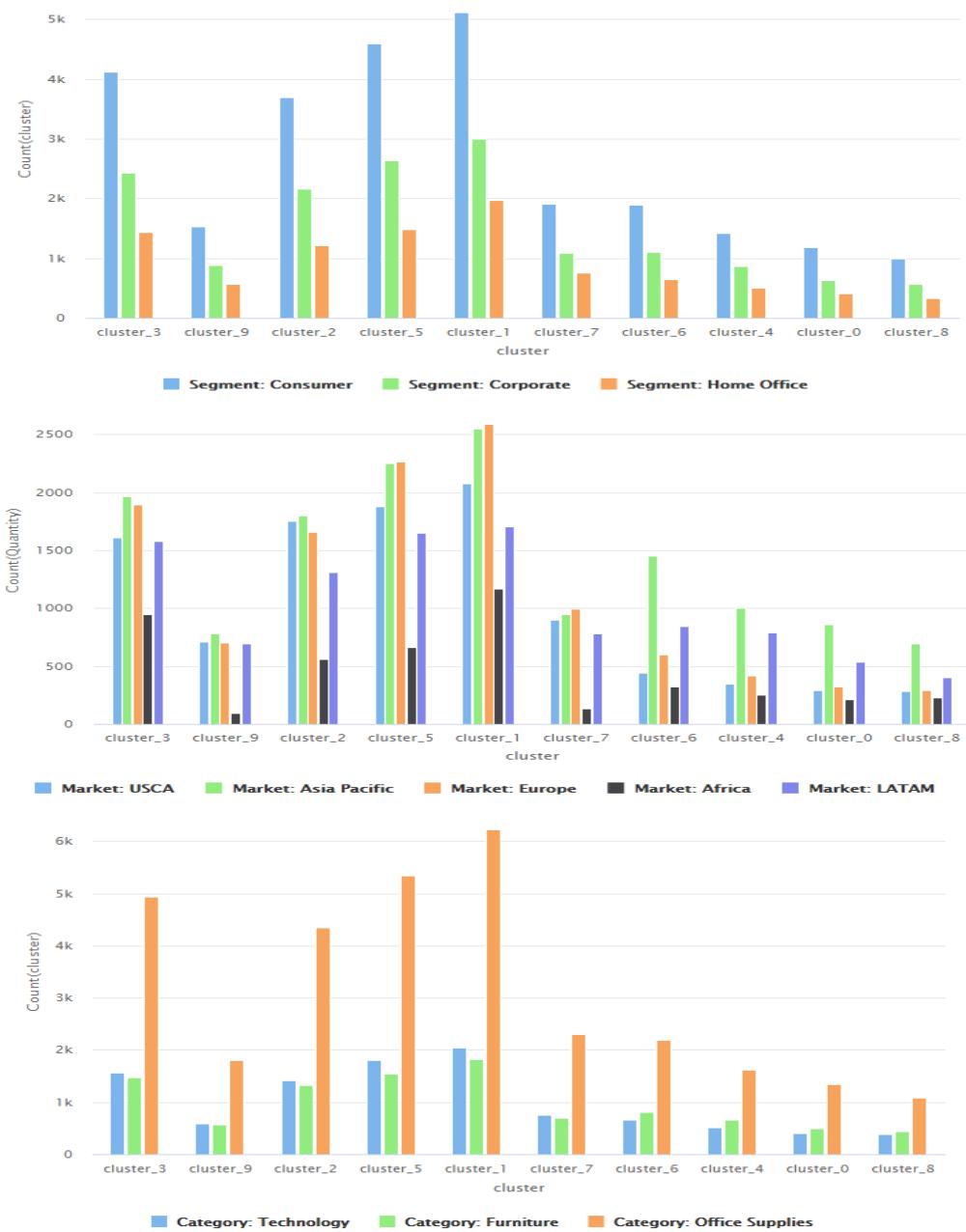
شکل ۱۱۲ – مقایسه خوشه ها در هر بخش مشتری

۴- مقایسه نتایج دو مدل

گام ۴ و نهایی، مقایسه نتایج حاصل از خوشه بندی به روش k -میانگین و خوشه بندی با ترکیب SOM و K-Mیانگین است. برای آن که امکان مقایسه وجود داشته باشد، عملگر Set Role به طور موقت به مدل گام ۲ افزوده و ویژگی های بازار و دسته بندی محصول و بخش مشتری به عنوان برچسب انتخاب شده اند تا بتوان توزیع آن ها بر اساس خوشه ها(مطابق گام ۳) نمایش داد.



شکل ۱۱۳- جایگاه عملگر Set Role

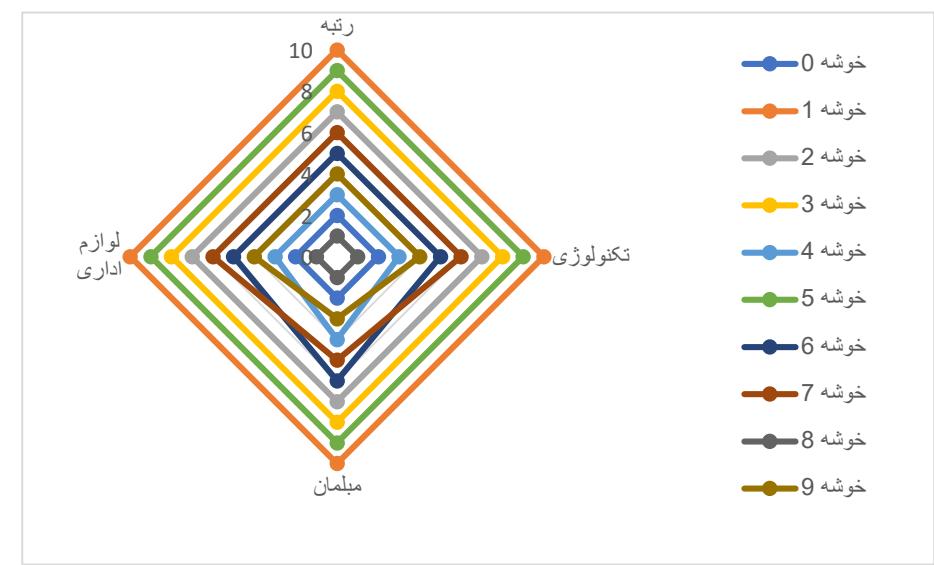
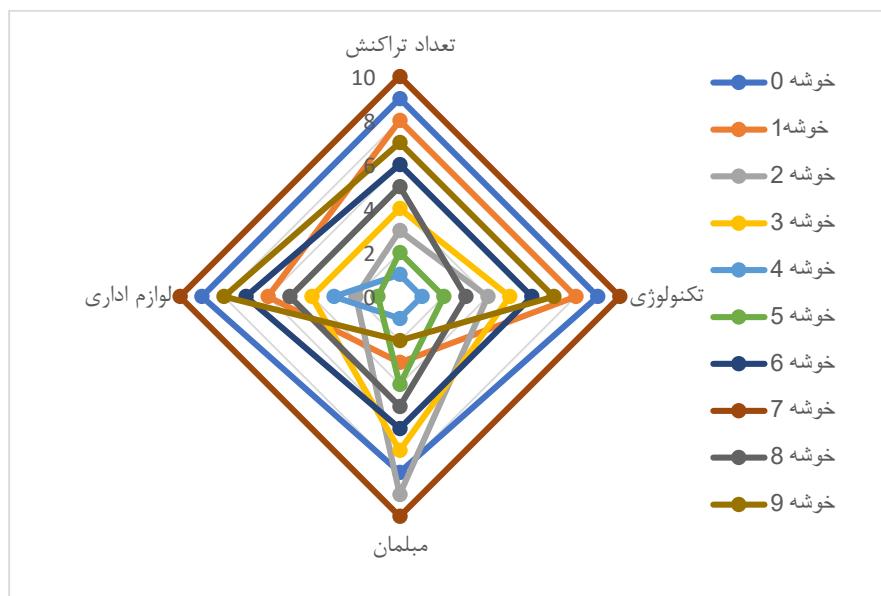
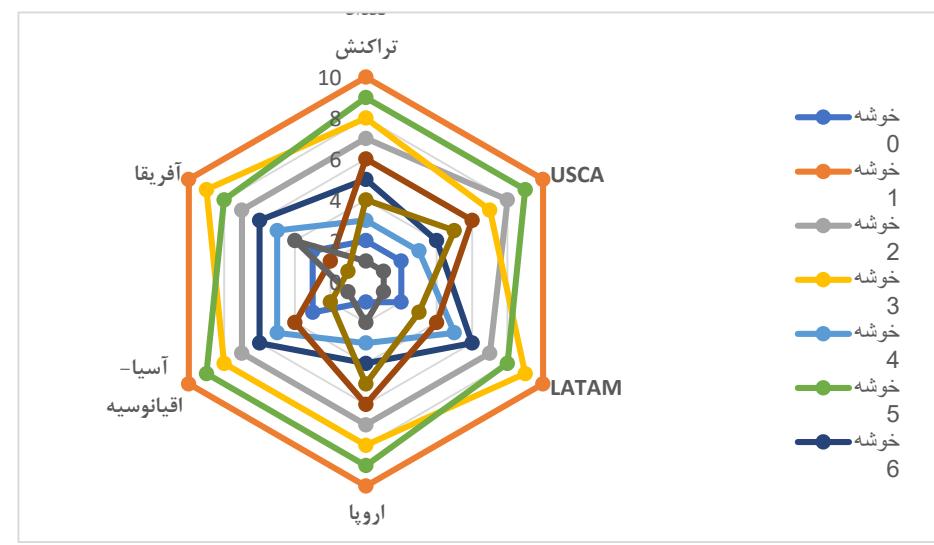
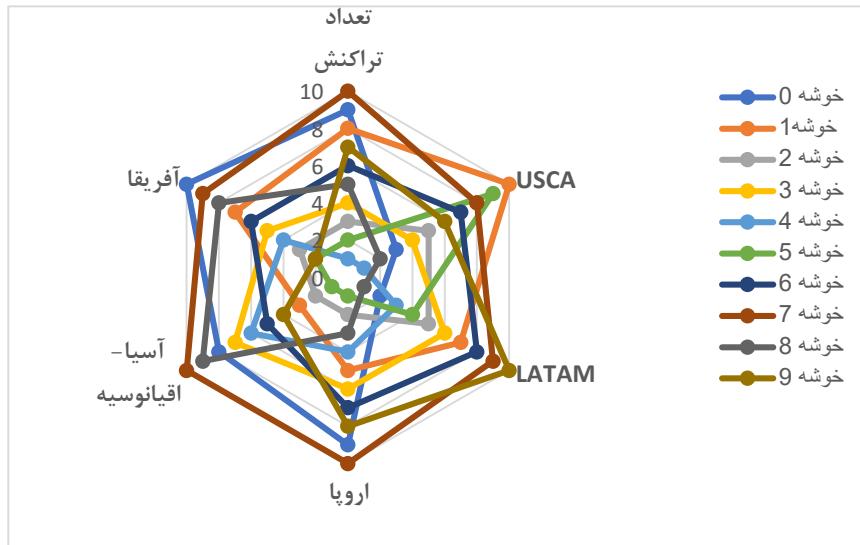


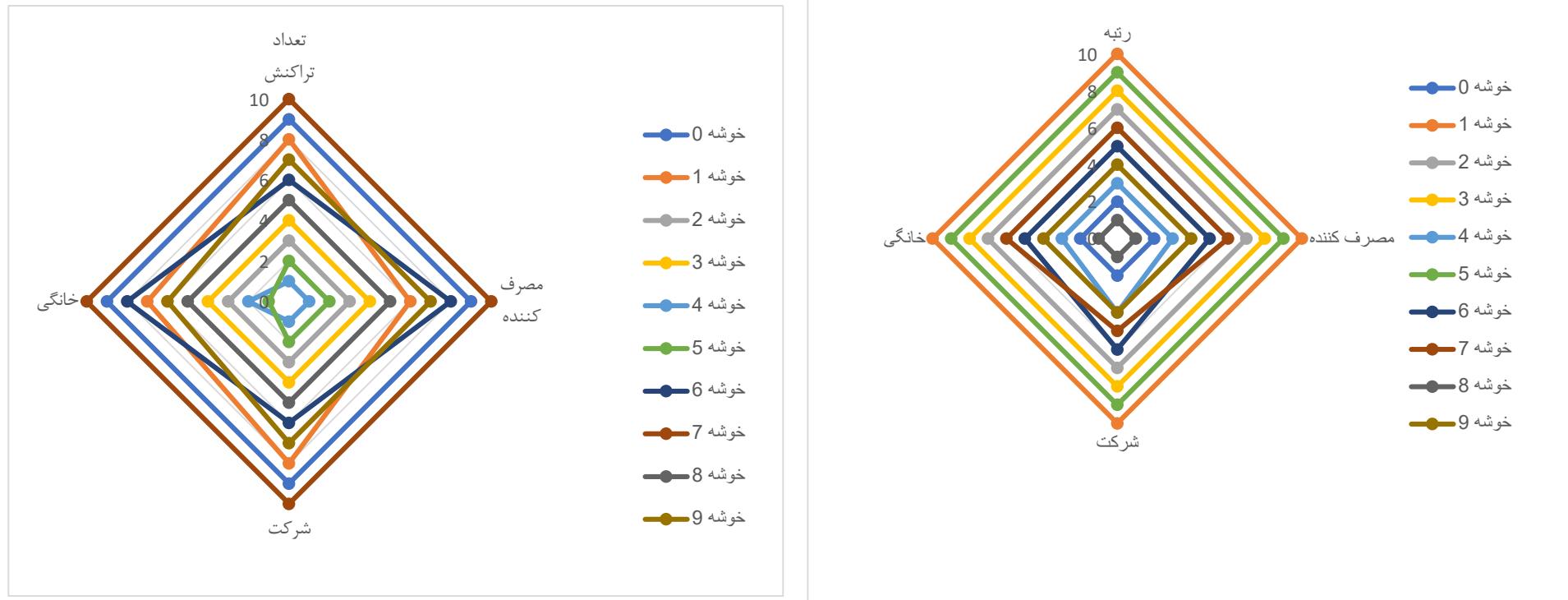
شکل ۱۱۴- توزیع متغیرهای کیفی اسمی در خوشه ها(مدل گام ۲)

جدول ۱۱- رتبه بندی خوشه ها بر اساس مشخصه های مختلف

رتبه در بخش خانگی	رتبه در بخش شرکت	رتبه در بخش صرف کننده	رتبه در دسته لوازم اداری	رتبه در دسته بندی مبلمان	رتبه در دسته بندی تکنولوژی	رتبه در بازار آفریقا	رتبه در بازار آسیا اقیانوسیه	رتبه در بازار اروپا	رتبه در بازار LATAM	رتبه در بازار USCA	رتبه تعداد تراکنش ها	نام خوشه
۹	۹	۹	۹	۹	۹	۸	۸	۱۰	۹	۹	۹	۰ خوشه
۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱ خوشه ۱
۴	۴	۴	۴	۴	۴	۴	۴	۴	۴	۳	۴	۲ خوشه
۳	۳	۳	۳	۳	۳	۲	۳	۳	۲	۴	۳	۳ خوشه ۳
۸	۸	۸	۸	۷	۸	۶	۶	۸	۶	۸	۸	۴ خوشه
۲	۲	۲	۲	۲	۲	۳	۲	۲	۳	۲	۲	۵ خوشه
۶	۶	۶	۶	۵	۶	۵	۵	۷	۵	۷	۶	۶ خوشه
۵	۵	۵	۵	۶	۵	۹	۷	۵	۷	۵	۵	۷ خوشه
۱۰	۱۰	۱۰	۱۰	۱۰	۱۰	۷	۱۰	۹	۱۰	۱۰	۱۰	۸ خوشه
۷	۷	۷	۷	۸	۷	۱۰	۹	۶	۸	۶	۷	۹ خوشه

این جدول برای مدل گام ۲ ترسیم شده است. نمودارهای راداری نیز در صفحه بعد ترسیم شده اند. نمودارهای سمت راست مربوط به مدل گام ۲ و نمودارهای سمت چپ مربوط به مدل گام ۳ می باشند.





شکل ۱۱۵ – مقایسه نتایج حاصل از گام ۲ و ۳

نمودارها شبیه به هم به نظر می‌رسند اما برای اطمینان، فاصله رتبه‌های خوشی مشابه (از نظر بزرگی خوشی) در هر دو مدل گام ۳ و ۴ محاسبه شده و عدد ۰،۰۰۹۰ حاصل آن است (پیوست‌ها). این عدد کوچک نشان از نزدیکی نتایج دو مدل دارد اما در گام ۳ به دلیل اعمال فرآیند کاهش ابعاد پیش از خوشی بندی و حذف مشخصه‌هایی که وزن کمی داشتند نتایج دقیق‌تر محاسبه شده است.

در نهایت، نتایج هر دو مدل قابل اکتفا و تقریباً یکسان می‌باشند. بخش بندی مشتریان بر مبنای نتایج به دست آمده نشان می‌دهد خوشی‌های مشتریان از نظر متوسط سودآوری در فاصله‌ای نزدیک قرار دارند. همچنین میان خوشی‌های مختلف در دسته بندی‌های مختلف محصول و بخش‌های مختلف مشتریان تفاوت کمی وجود دارد، به این معنا که برای مثال اگر خوشی‌ای دومین خوشی دسته لوازم اداری است به احتمال زیاد دومین خوشی مبلغان نیز هست. این موضوع برای بازارها متفاوت است و خوشی‌ها نسبت به

بازار های مختلف رفتارهای متفاوتی دارند که هر کدام در بخش نتایج گام های ۲ و ۳ به تفصیل بیان شدند. نتایج تفاوت هایی نیز با یکدیگر دارند. در خوشه بندی گام ۲، یک دسته‌ی بزرگ وجود دارد که در همه بازارها کنترل را در اختیار دارد. این موضوع به دلیل توزیع بیشتر تراکنش ها در این دسته و در نتیجه غلبه این دسته در همه مشخصه ها دارد اما نتیجه‌ی خوشه بندی با SOM متفاوت و توزیع داده ها در خوشه ها متناسب تر است. خلاصه تفاوت ها به صورت زیر است:

جدول ۱۲ – تفاوت نتایج مدل ها

خوشه بندی به وسیله SOM و K-میانگین	خوشه بندی به وسیله K-میانگین
توزیع متناسب داده ها در خوشه ها	توزیع داده ها در خوشه ها کمی نامتناسب
خوشه بزرگ، خوشه اول بازارهای آمریکای شمالی، آمریکای جنوبی و آفریقا نیست(با اختلاف کم).	خوشه بزرگ، خوشه اول همه بازارها است.
در ۳ خوشه رفتار متفاوت(خوشه ۵ و ۳ و ۲)	رفتار همه خوشه ها در همه دسته محصولات یکسان

تصویرسازی خوشه بندی به کمک SOM بهتر و قابل درک تر بود و در خوشه بندی گام ۲، نمایش رکورد ها کمی گیج کننده و توزیع بسیاری از داده ها درهم و تشخیص خوشه آن ها دشوار بود. همچنین در انتهای مدل گام ۳ می‌توان دید که عملگر تصویرسازی SOM وجود دارد که با اجرای آن می‌توان توزیع مشخصه اول و دوم حاصل از SOM و همچنین مشخصه‌ی برچسب یا label را میان شبکه عصبی SOM مشاهده کرد که این مورد نیز برتری دیگری از مدل گام ۲ است.

پیوست ها

این بخش راهنمای فایل های پیوست این پژوهش می باشد.

جدول ۱۳- راهنمای فایل های پیوست

نام فایل	فرمت	توضیحات
global_superstore_2016	xlsx	دیتابیس اولیه
edited - global_superstore_2016-sheet 1	csv	صفحه (sheet) اول دیتابیس – ورودی نرم افزار RapidMiner
edited - global_superstore_2016-sheet 2	csv	صفحه (sheet) دوم دیتابیس – ورودی نرم افزار RapidMiner
11	py	سورس کد پایتون برای محاسبه تمایل داده ها به خوش بندی
Edited_global_superstore_2016	xlsx	فایل اکسل ویرایش شده مناسب پایتون (فایل "11")
Centroids	xlsx	مراکز خوش ها و محاسبه فاصله بین رتبه خوش های گام ۲ و ۳

