گزارش تکلیف سوم درس الگوریتم های علوم داده

نام و نام خانوادگی: فاطمه ترودی

شماره دانشجویی: ۴۰۳۴۲۲۰۴۸

نام استاد: دکتر سعیدرضا خردپیشه

نیمسال دوم ۰۴-۱۴۰۳

فهرست مطالب

۱. پیش‌پردازش داده‌ها و تحلیل اکتشافی

۱/۱. خلاصه پیش‌پردازش داده‌ها

در این بخش، فرآیند پیش‌پردازش و ادغام داده‌های مجموعه داده‌های Olist برای ایجاد یک مجموعه داده در سطح مشتری انجام شد. جدول‌های مرتبط شامل customers_df ، orders_df، payments_df، reviews_df، items_df، products_df و category_translation_df انتخاب شدند. جدول‌های geolocation_df و sellers_df به دلیل عدم ارتباط مستقیم با رفتار مشتری حذف شدند، زیرا اطلاعات جغرافیایی دقیق و داده‌های فروشندگان برای بخش‌بندی مشتریان ضروری نبودند. فرآیند ادغام با استفاده از کلیدهای موجود در نمودار رابطه‌ای) مانند customer_id و order_id انجام شد تا تمام مشتریان حفظ شوند.

۱/۲. مدیریت مقادیر گم‌شده

در این بخش، فرآیند پیش‌پردازش و ادغام داده‌های مجموعه داده‌های Olist برای ایجاد یک مجموعه داده در سطح مشتری انجام شد. جدول‌های مرتبط شامل customers_df ، orders_df، payments_df، reviews_df، items_df، products_df و category_translation_df انتخاب شدند. جدول‌های geolocation_df و sellers_df به دلیل عدم ارتباط مستقیم با رفتار مشتری حذف شدند، زیرا اطلاعات جغرافیایی دقیق و داده‌های فروشندگان برای بخش‌بندی مشتریان ضروری نبودند. فرآیند ادغام با استفاده از کلیدهای موجود در نمودار رابطه‌ای) مانند customer_id و order_id انجام شد تا تمام مشتریان حفظ شوند.

پیش از ادغام: مقادیر گمشده در جدول‌های جداگانه بر اساس اطلاعات اولیه مدیریت شدند:

- در orders_df، تعداد ۱۶۰ تا ۲۹۶۵ مقدار گمشده در تاریخ‌های تحویل با تاریخ تخمینی تحویل (order_estimated_delivery_date) پر شدند.
- در products_df، ۶۱۰ مقدار گمشده در product_category_name با "unknown" و مقادیر عددی مرتبط (مانند طول نام محصول) با میانگین پر شدند. همچنین، ۲ ردیف با ابعاد فیزیکی گمشده حذف شدند.

پس از ادغام: مقادیر گمشده جدید در جدول ادغام‌شده orders_full_df شناسایی و مدیریت شدند:

- ۳ مقدار گمشده در payment_sequential، payment_installments و payment_value با صفر و در payment_type با حالت (mode) پر شدند.
- ۸۶۱ مقدار گمشده در review_score با میانگین پر شدند.
- ۲۰ مقدار گمشده در product_category_name_english با "unknown" پر شدند.

پس از این مراحل، هیچ مقدار گمشده‌ای در مجموعه داده باقی نماند.

۱/۳. مدیریت مقادیر پرت و انواع داده

مقادیر پرت در ویژگی‌های عددی مانند total_spend، avg_spend، total_price، total_freight، delivery_time و recency در صدک ۹۹ محدود شدند تا از تأثیر مقادیر افراطی بر تحلیل جلوگیری شود.

تاریخ‌ها به فرمت datetime تبدیل شدند (مانند order_purchase_timestamp)، و متغیرهای رسته‌ای (categorical) با کدگذاری one-hot encoding برای سازگاری با الگوریتم‌های خوشه‌بندی آماده شدند.

۱/۴. تجمیع داده‌ها در سطح مشتری

داده‌های ادغام‌شده بر اساس customer_unique_id تجمیع شدند تا هر ردیف نماینده یک مشتری باشد. ویژگی‌های زیر محاسبه شدند:

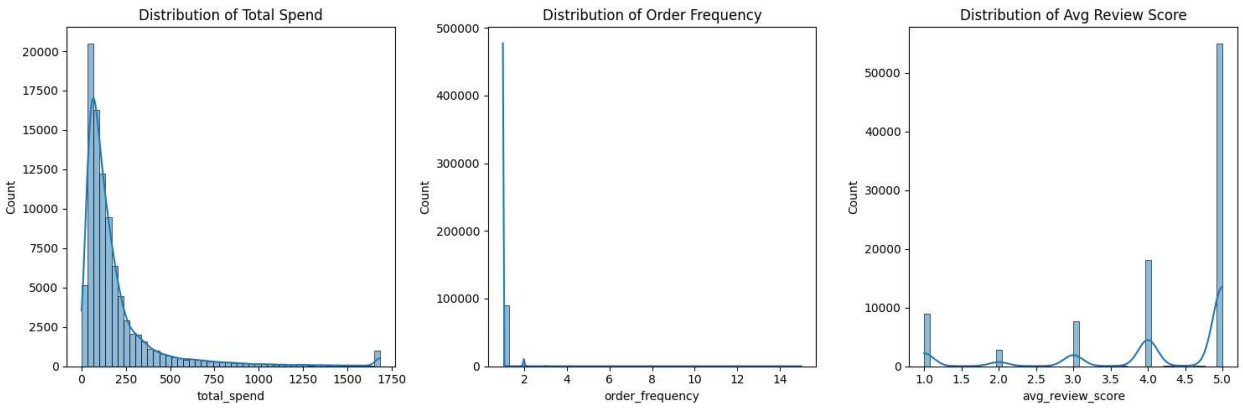
- **total_spend و avg_spend**: مجموع و میانگین هزینه پرداخت‌ها برای شناسایی قدرت خرید.
- **order_frequency**: تعداد سفارش‌های منحصربه‌فرد برای اندازه‌گیری وفاداری.
- **avg_review_score**:میانگین امتیاز بررسی‌ها برای سنجش رضایت مشتری.
- **total_price و total_freight**: مجموع قیمت اقلام و هزینه حمل‌ونقل برای جزئیات هزینه.
- **recency**: تعداد روزهای گذشته از آخرین خرید.
- **delivery_time**: میانگین زمان تحویل برای تجربه لجستیک.
- **customer_state**: ایالت مشتری برای تحلیل جغرافیایی.
- **top_category**: دسته‌بندی محبوب محصولات برای ترجیحات خرید.
- **top_payment_type**: نوع پرداخت محبوب برای رفتار پرداخت.

مقادیر گمشده پس از تجمیع با صفر (برای هزینه‌ها) یا میانگین (برای امتیازها و زمان تحویل) پر شدند تا مجموعه داده کامل شود.

۱٫۵. تحلیل اکتشافی داده‌ها

توزیع ویژگی‌های عددی

نمودارهای هیستوگرام برای total_spend، order_frequency و avg_review_score ترسیم شدند.

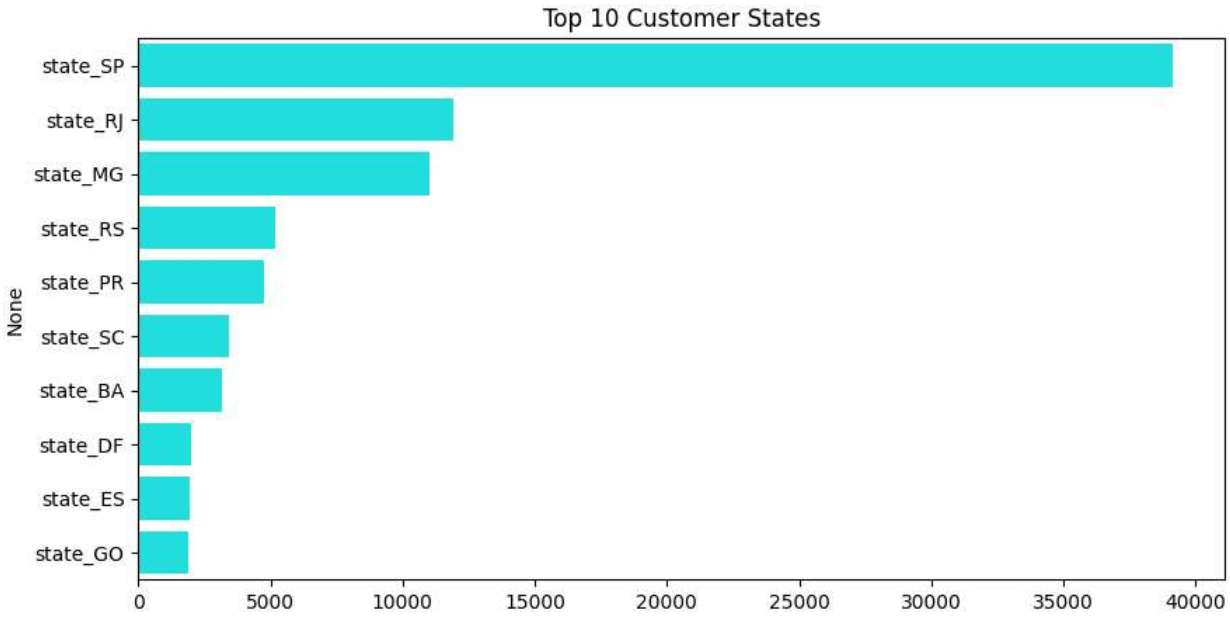


شکل ۱- نمودار هیستوگرام برای متغیرهای total_spend، order_frequency و avg_review_score

مشاهده می‌شود که هر سه متغیر دارای توزیع نرمال نیستند و برای مثال متغیر total_spend دارای چولگی به سمت راست می‌باشد.

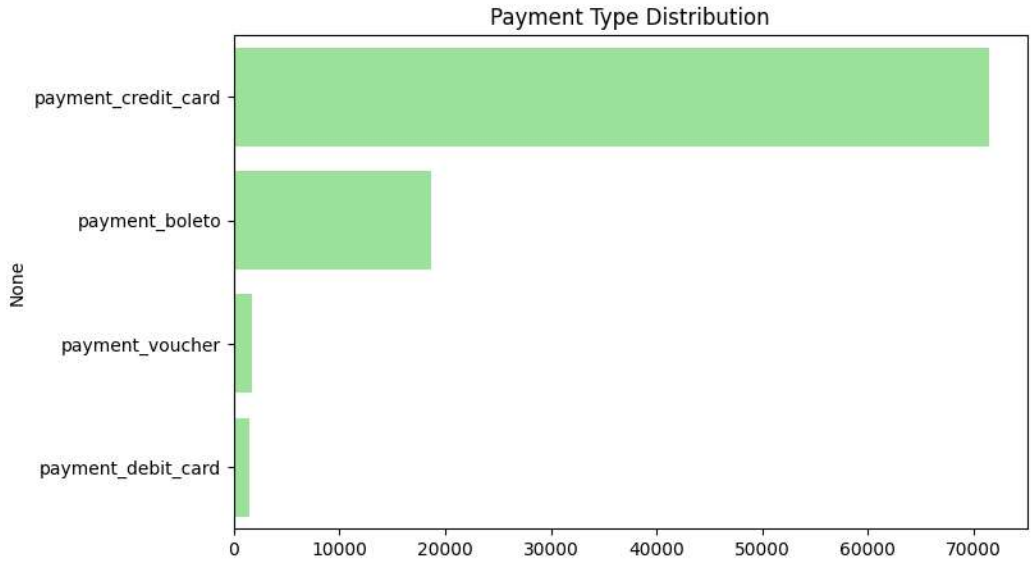
توزیع جغرافیایی

نمودار میله‌ای برای ۱۰ ایالت برتر ترسیم شدند که به صورت زیر می‌باشد:



شکل ۲- نمودار میله‌ای برای ۱۰ ایالت برتر

توزیع نوع پرداخت

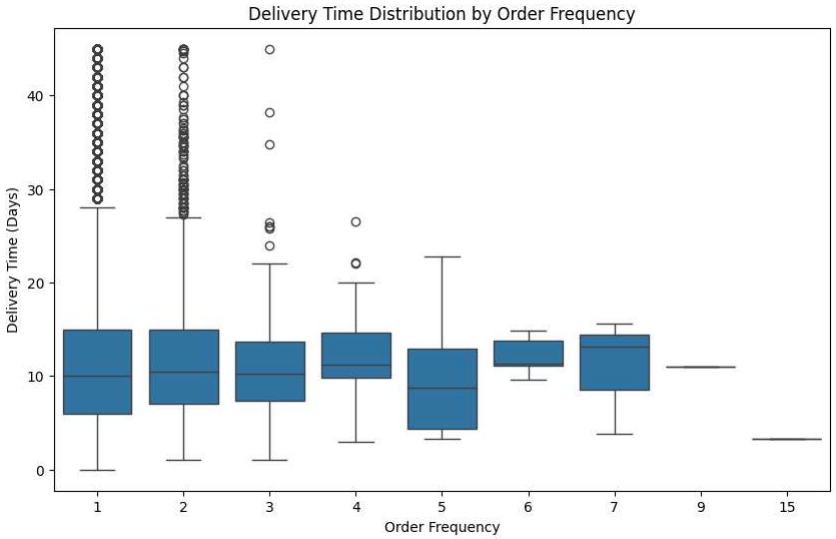


شکل ۳- نمودار میله‌ای توزیع نوع پرداخت

مشاهده می‌شود که پرداخت با کارت اعتباری رایج‌ترین روش پرداخت می‌باشد.

زمان تحویل بر اساس فرکانس سفارش

نمودار جعبه‌ای (box plot) برای delivery_time بر اساس order_frequency ترسیم شد تا بررسی شود آیا مشتریان مکرر زمان تحویل متفاوتی دارند.

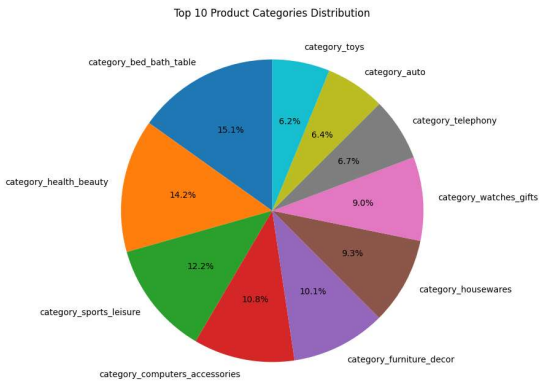


شکل ۴- نمودار جعبه‌ای برای زمان تحویل بر اساس فرکانس سفارش

الگوی خاصی در نمودار مشاهده نمی‌شود و برای هر فرکانس سفارش، طول زمان تحویل (تعداد روز) تفاوت آنچنانی نمی‌کند.

توزیع دسته‌بندی برتر محصولات

نمودار دایره‌ای (pie chart) برای ۱۰ دسته‌بندی برتر محصولات نشان‌دهنده ترجیحات خرید مشتریان است.



شکل ۵- نمودار جعبه‌ای برای زمان تحویل بر اساس فرکانس سفارش

خلاصه آمار توصیفی برای متغیرهای عددی

Statistical Summary of Numerical Features:

	total_spend	order_frequency	avg_review_score	recency	delivery_time
count	93358.000000	93358.000000	93358.000000	93358.000000	93358.000000
mean	193.443005	1.033420	4.159056	2687.100388	11.930351
std	256.316755	0.209097	1.277958	151.630451	8.378599
min	0.000000	1.000000	1.000000	2450.000000	0.000000
25%	63.830000	1.000000	4.000000	2564.000000	6.000000
50%	113.140000	1.000000	5.000000	2668.000000	10.000000
75%	202.637500	1.000000	5.000000	2796.000000	15.000000
max	1684.274800	15.000000	5.000000	3025.000000	45.000000

شکل ۶- آمار توصیفی برای متغیرهای عددی

۲. خوشه‌بندی پایه (Baseline Clustering)

۲٫۱. آماده‌سازی داده‌ها

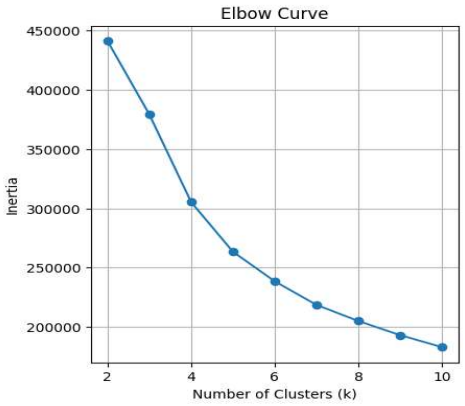
- داده‌ها از مجموعه‌های مختلف شامل اطلاعات مشتریان، سفارش‌ها، پرداخت‌ها، بررسی‌ها، اقلام سفارش‌ها، محصولات و ترجمه دسته‌بندی محصولات جمع‌آوری و ادغام شدند.
- داده‌های ناقص (مانند زمان تأیید سفارش، تاریخ تحویل و امتیاز بررسی) با استفاده از مقادیر پیش‌فرض (مانند میانگین یا مُد) پر شدند.
- ویژگی‌های سطح مشتری با تجمیع داده‌ها در سطح شناسه منحصربه‌فرد مشتری (Customer-level Aggregation) ایجاد شد. این ویژگی‌ها شامل مجموع هزینه (total_spend)، میانگین هزینه (avg_spend)، فراوانی سفارش order_frequency، میانگین امتیاز بررسی avg_review_score، تازگی (recency)، و میانگین زمان تحویل (delivery_time) بودند.
- برای مدیریت مقادیر پرت، از روش چارک ۹۹ درصد استفاده شد و سپس تبدیل لگاریتمی برای کاهش ناهمگونی در توزیع داده‌ها اعمال گردید.
- در ابتدا، ویژگی‌های رسته‌ای (مانند ایالت مشتری، دسته‌بندی محصولات اصلی، و نوع پرداخت اصلی) با One-Hot Encoding رمزگشایی شدند، اما در مراحل بعدی به دلیل افزایش نویز، این ویژگی‌ها حذف شدند و فقط ویژگی‌های عددی مورد استفاده قرار گرفتند.

۲/۲. پیش‌پردازش و مقیاس‌بندی

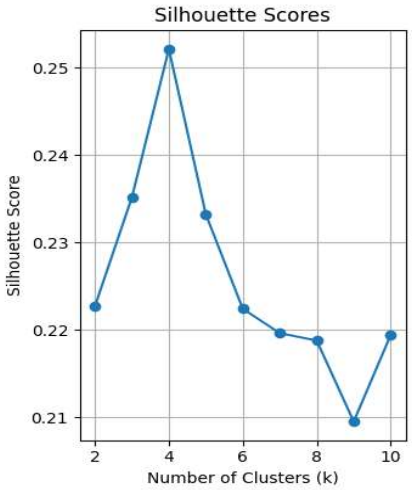
از مقیاس‌کننده استاندارد (StandardScaler) برای نرمال‌سازی داده‌ها با میانگین صفر و انحراف معیار واحد استفاده شد. این روش به جای مقیاس‌کننده MinMax انتخاب شد تا تأثیر بهتری بر داده‌های لگاریتمی‌شده داشته باشد.

۲/۳. خوشه‌بندی با K-Means

الگوریتم K-means با تعداد خوشه‌های مختلف (از ۲ تا ۱۰) آزمایش شد. برای انتخاب تعداد بهینه خوشه‌ها، از روش‌های منحنی آرنج (Elbow Method) و Silhouette Score استفاده شد که نمودار های آن به شکل زیر می‌باشد:



شکل ۷- نمودار منحنی Elbow برای انتخاب تعداد بهینه خوشه‌ها



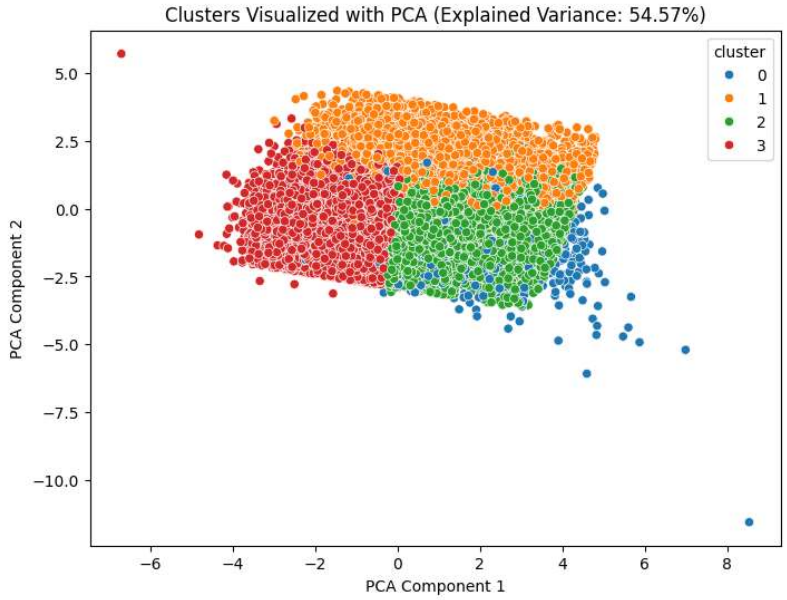
شکل ۸- نمودار Silhouette Score برای انتخاب تعداد بهینه خوشه‌ها

تعداد بهینه خوشه‌ها با توجه به نمودار Elbow و Silhouette Score، ۴ (k=4) در نظر گرفته شد و خوشه‌بندی با الگوریتم K-Means اجرا شد.

۲/۴. نتایج

Silhouette Score برای k=4 به ۰.۲۵۲ رسید که نشان‌دهنده جداسازی خوب خوشه‌ها است (بالاتر از آستانه ۰.۲-۰.۳). امتیاز Davies-Bouldin به ۱.۲۱۷ کاهش یافت که نشان‌دهنده فشردگی مناسب خوشه‌هاست (کمتر از آستانه ۱.۵).

تجسم خوشه‌ها با استفاده از تحلیل مؤلفه‌های اصلی (PCA) انجام شد که نشان داد ۵۴.۵۷٪ از واریانس داده‌ها با دو مؤلفه توضیح داده می‌شود. این تجسم چهار خوشه را با جداسازی نسبی (هرچند با برخی همپوشانی) نشان داد که نمودار آن به صورت زیر می‌باشد:



شکل ۹- نمودار خوشه‌بندی K-Means رسم‌شده با استفاده از PCA

حذف ویژگی‌های رسته‌ای و تمرکز بر ویژگی‌های عددی پس از تبدیل لگاریتمی، بهبود قابل‌توجهی در کیفیت خوشه‌بندی ایجاد کرد. Silhouette Score از ۰.۱۳۴ (k=2) به ۰.۲۵۲ (k=4) و واریانس توضیح‌داده‌شده توسط PCA از ۸.۹۳٪ به ۵۴.۵۷٪ افزایش یافت.

چهار خوشه شناسایی‌شده با استفاده از ویژگی‌های مالی (مانند total_spend)، رفتاری (مانند order_frequency)، و زمانی (مانند recency و delivery_time) به دست آمدند که نشان‌دهنده پتانسیل تقسیم‌بندی مشتریان به گروه‌های معنادار است.

در این بخش، با بهینه‌سازی الگوریتم K-means و تمرکز بر ویژگی‌های عددی، توانستیم مدل خوشه‌بندی پایه‌ای با عملکرد خوب (Silhouette Score برابر با ۰.۲۵۲) ایجاد کنیم.

۳. انواع مدل‌ها و کاهش بعد

روش های خوشه‌بندی

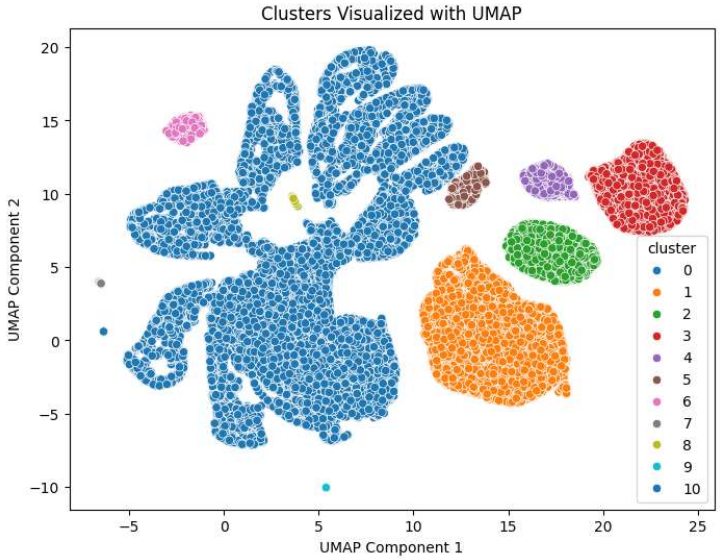
ما سه روش خوشه‌بندی را آزمایش کردیم:

۱. خوشه‌بندی K-Means (از روش های Partitioning)

- **پارامترها:** ۴ خوشه، فاصله اقلیدسی
- **داده‌ها**: کل مجموعه داده
- **کاهش ابعاد:** PCA (۵۴.۵۷ درصد واریانس توضیح داده‌شده)
- **نتایج**:
- Silhouette Score: ۰.۲۵۲
- امتیاز Davies-Bouldin: ۱.۲۱۷
- **تجسم**: نمودار PCA چهار خوشه با جداسازی متوسط اما با مقداری همپوشانی نشان داده شد که نمودار آن را در بخش‌های قبلی مشاهده کردیم.

۲. خوشه‌بندی DBSCAN (از روش های مبتنی بر چگالی)

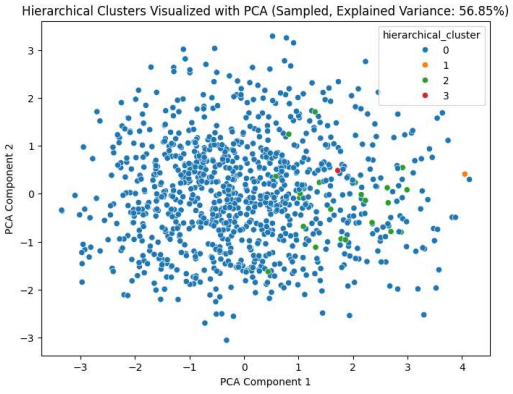
- **پارامترها:** eps=0.4، min_samples=8
- **داده‌ها** :کل مجموعه داده
- **کاهش ابعاد**: UMAP
- **نتایج:**
- تعداد خوشه‌ها: ۱۱
- نقاط نویز: ۰
- Silhouette Score: ۰.۰۲۴
- امتیاز Davies-Bouldin: ۰.۸۰
- **تجسم**:



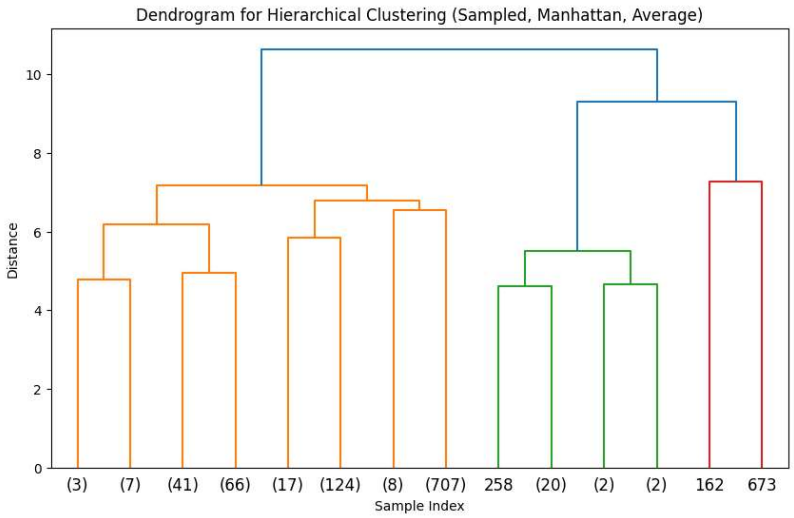
شکل ۱۰- نمودار خوشه‌بندی DBSCAN رسم‌شده با استفاده از UMAP

۳. خوشه‌بندی سلسله مراتبی (Agglomerative Hierarchical Clustering)

- **پارامترها:** ۴ خوشه، فاصله Manhattan، پیوند میانگین (Average Linkage)
- **داده‌ها**: نمونه‌ای از ۱۰۰۰ مشتری (به دلیل محدودیت حافظه).
- **کاهش ابعاد:** PCA (حدود ۵۶.۸۵ درصد واریانس توضیح داده‌شده)
- **نتایج**:
  - Silhouette Score: ۰.۴۶۲
  - امتیاز Davies-Bouldin: ۰.۵۱۴
- **تجسم:** نمودار PCA خوشه‌های جدا شده و دندروگرام ساختار سلسله‌مراتبی را با جداسازی واضح در سطح ۴ خوشه به صورت زیر نشان می‌دهند:



شکل ۱۱- نمودار خوشه‌بندی سلسله مراتبی رسم‌شده با استفاده از PCA



شکل ۱۲- نمودار Dendrogram خوشه‌بندی سلسله مراتبی

مزایا و معایب

پایداری:

- **K-means**: بسیار پایدار (نتایج ثابت با ۴ خوشه).
- **DBSCAN**: پایداری کم (حساس به پارامترها، تعداد خوشه‌ها متغیر بود).
- **سلسله‌مراتبی**: پایداری متوسط (نتایج برای نمونه ثابت، اما نمونه‌برداری باعث تغییرپذیری می‌شود).

تفسیرپذیری:

- **K-means**: ۴ خوشه با جداسازی متوسط، تفسیر آسان.
- **DBSCAN**: ۱۱ خوشه با همپوشانی زیاد، تفسیر دشوار.
- **سلسله‌مراتبی**: ۴ خوشه با جداسازی خوب، دندروگرام و PCA تفسیر را تسهیل می‌کنند.

کاربرد:

- **K-means**: خوشه‌ها برای بازاریابی قابل استفاده‌اند، اما همپوشانی ممکن است دقت را کاهش دهد.
- **DBSCAN**: به دلیل همپوشانی خوشه‌ها، کاربرد محدودی دارد.
- **سلسله‌مراتبی**: خوشه‌های باکیفیت برای بازاریابی مناسب‌اند، اما نمونه‌برداری کاربرد آن را محدود می‌کند.

کاهش ابعاد:

- **PCA**: واریانس قابل اندازه‌گیری، اما ممکن است الگوهای غیرخطی را از دست بدهد.
- **UMAP**: الگوهای غیرخطی را بهتر نشان می‌دهد، اما واریانس قابل اندازه‌گیری ندارد.

۴. پروفایل‌بندی خوشه‌ها و بینش‌های تجاری

در این بخش، هدف ما تحلیل چهار خوشه نهایی که با استفاده از الگوریتم K-means روی داده‌ها شناسایی شده‌اند، بود. این خوشه‌ها بر اساس ویژگی‌هایی مانند میانگین ارزش سفارش، تعداد سفارش‌ها، امتیاز بررسی (review score)، تازگی خرید (recency)، و زمان تحویل شکل گرفته‌اند. برای هر خوشه، آمار کلیدی محاسبه شد، برچسب بازاریابی (marketing label) تعیین گردید، و حداقل دو پیشنهاد عملی ارائه شد.

پروفایل‌بندی خوشه‌ها

بر اساس آمارهای کلیدی بدست آمده چهار خوشه به شرح زیر پروفایل‌بندی شدند:

خوشه صفر

- **آمار کلیدی**:
  - میانگین ارزش سفارش (BRL): ۲۵۹.۴۰
  - میانگین تعداد سفارش: ۱.۰
  - میانگین امتیاز بررسی: ۴.۶۵ (نگرش مثبت)
  - میانگین recency (روز): ۲۶۸۷.۲۷ (غیرفعال به مدت ۷.۴ سال)
  - میانگین زمان تحویل (روز): ۱۱.۸۱
  - منطقه برتر: SP (سائو پائولو)
  - دسته‌بندی (category) برتر: health_beauty
  - میانگین مجموع هزینه: ۳۲۴.۹۵
- **برچسب بازاریابی:** علاقمندان غیرفعال به سلامت
- **توصیه‌های عملی**:

۱. اجرای کمپین تخفیف هدفمند (مثلاً ۲۰٪ تخفیف روی محصولات health_beauty) با ایمیل‌های شخصی‌سازی‌شده برای بازگرداندن این مشتریان غیرفعال با ارزش بالا.
۲. ارائه پاداش وفاداری (مثلاً حمل رایگان در خرید بعدی) برای تشویق بازگشت آن‌ها.

خوشه ۱

- **آمار کلیدی**:
  - میانگین ارزش سفارش: ۶۴.۵۹
  - میانگین تعداد سفارش: ۱.۰۰
  - میانگین امتیاز بررسی: ۴.۶۱ (نگرش مثبت)
  - میانگین recency (روز): ۲۶۸۶.۴۵ (غیرفعال به مدت ~۷.۴ سال)
  - میانگین زمان تحویل (روز): ۹.۴۸
  - منطقه برتر: SP (سائو پائولو)
  - دسته‌بندی برتر: health_beauty
  - میانگین مجموع هزینه(BRL): 68.13
- **برچسب بازاریابی**: خریداران غیرفعال با بودجه کم
- **توصیه‌های عملی**:

۱. اجرای کمپین بازگردانی با بسته‌های ارزان‌قیمت health_beauty برای جذب این مشتریان کم‌خرج و غیرفعال

۲. بهینه‌سازی مسیرهای تحویل در سائو پائولو برای کاهش زمان تحویل (زیر ۹.۴۸ روز) به بهبود تجربه آن‌ها پس از بازگشت.

خوشه ۲

- **آمار کلیدی**:
  - میانگین ارزش سفارش: ۱۴۵.۲۰
  - میانگین تعداد سفارش: ۲.۱۱
  - میانگین امتیاز بررسی: ۴.۲۱ (نگرش مثبت)
  - میانگین recency (روز): ۲۶۶۹.۵۸ (غیرفعال به مدت ~۷.۳ سال)
  - میانگین زمان تحویل (روز): ۱۱.۸۱
  - منطقه برتر: SP (سائو پائولو)
  - دسته‌بندی برتر: bed_bath_table
  - میانگین مجموع هزینه: ۴۰۳.۵۸
- **برچسب بازاریابی:** عاشقان وفادار دکوراسیون خانه
- **توصیه‌های عملی:**

۱. ارائه تخفیف فصلی روی محصولات bed_bath_table (مثلاً ۱۵٪ در تعطیلات) برای استفاده از تعداد سفارش‌های بالاتر.
۲. راه‌اندازی برنامه وفاداری مبتنی بر امتیاز که خریداران مکرر (مانند این خوشه) امتیاز دوبرابر برای خرید بعدی کسب کنند.

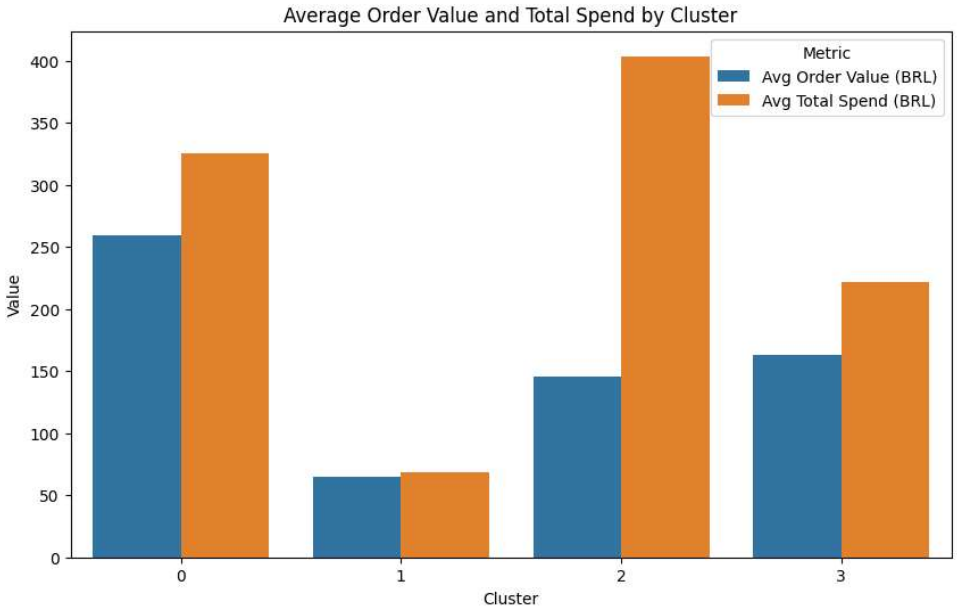
خوشه ۳

- **آمار کلیدی :**
  - میانگین ارزش سفارش: ۱۶۲.۷۳
  - میانگین تعداد سفارش: ۱.۰۰
  - میانگین امتیاز بررسی: ۱.۶۰ (نگرش منفی)
  - میانگین recency (روز): ۲۶۹۲.۲۱ (غیرفعال به مدت ~۷.۴ سال)
  - میانگین زمان تحویل (روز): ۱۹.۷۸
  - منطقه برتر: SP (سائو پائولو)
  - دسته‌بندی برتر: bed_bath_table
  - میانگین مجموع هزینه: ۲۲۱.۳۳
- **برچسب بازاریابی**: خریداران ناراضی با تحویل کند
- **توصیه‌های عملی:**

۱. بهبود کارایی تحویل در سائو پائولو با همکاری پیک‌های محلی سریع‌تر برای کاهش زمان تحویل (فعلی ۱۹.۷۸ روز) و بازسازی اعتماد.
۲. ارسال عذرخواهی شخصی‌سازی‌شده با تخفیف برای جبران نگرش منفی و تشویق خرید مجدد.

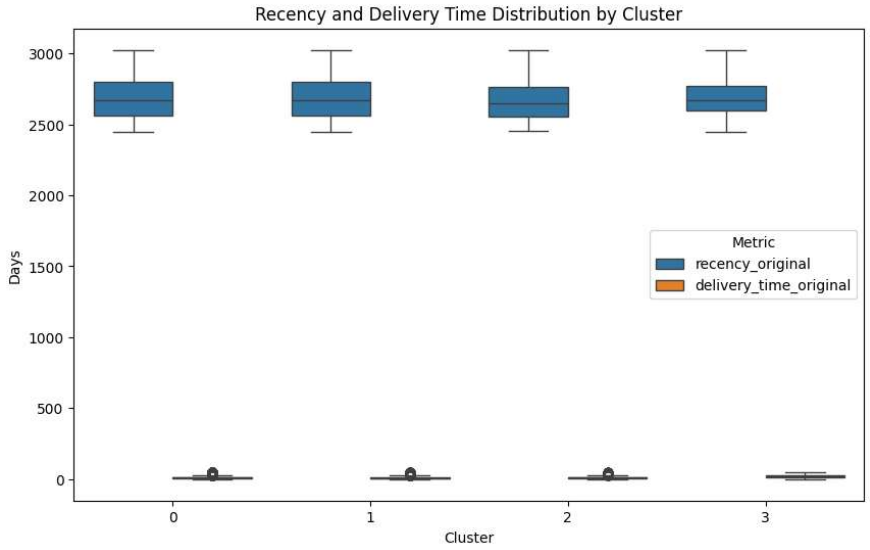
مشاهدات از نمودارها

- **میانگین ارزش سفارش و مجموع هزینه بر اساس خوشه:**
  - خوشه ۲ بالاترین میانگین مجموع هزینه و ارزش سفارش متوسط را دارد که نشان‌دهنده تعداد سفارش بالاتر (۲.۱۱) است.
  - خوشه ۰ ارزش سفارش بالا و مجموع هزینه را با وجود فرکانس ۱.۰۰ نشان می‌دهد.
  - خوشه‌های ۱ و ۳ ارزش کمتری دارند، با خوشه ۱ پایین‌ترین و خوشه ۳ متوسط.



شکل ۱۳- نمودار میانگین ارزش سفارش و مجموع هزینه بر اساس خوشه

- **توزیع recency و زمان تحویل بر اساس خوشه:**
  - همه خوشه‌ها تازگی بالایی (حدود ۲۵۰۰–۳۰۰۰ روز) دارند، که نشان‌دهنده غیرفعال بودن مشتریان به مدت ۷–۸ سال است، احتمالاً به دلیل قدیمی بودن داده‌ها یا نبود خرید اخیر.
  - زمان تحویل توزیع گسترده‌تری دارد، با خوشه ۳ طولانی‌ترین میانگین (۱۹.۷۸ روز) که نیاز به بهبود لجستیک را تأیید می‌کند.



شکل ۱۴- نمودار توزیع recency و زمان تحویل بر اساس خوشه