Proje Adı

Müşteri Segmentasyonu Uygulaması

İçindekiler

1. Giriş

Müşteri Segmentasyonu benzer davranışlar sergileyen müşterilerin ihtiyaç ve beklentilerine göre alt gruplara ayrılarak amaca uygun en doğru hedef kitleyi belirlemede kullanılan bir yöntemdir.

Proje kapsamında kişilerin harcama alışkanlıkları ile aile yapıları arasındaki bağıntı ortaya çıkarılmak üzere bir yemek firmasının müşteri verileri kullanılarak müşteri segmentasyonu gerçekleştirilmiştir.

Veri seti aşağıda sıralanmış sütunlardan oluşmaktadır:

- ID: Müşterinin benzersiz tanımlayıcısı
- Year_Birth: Müşterinin doğum yılı
- Education: Müşterinin eğitim seviyesi
- Marital_Status: Müşterinin medeni durumu
- Income: Müşterinin yıllık hane geliri
- Kidhome: Müşterinin evindeki çocuk sayısı
- Teenhome: Müşterinin evindeki gençlerin sayısı
- Dt_Customer: Müşterinin şirkete kayıt tarihi
- Recency: Müşterinin son satın alımından bu yana geçen gün sayısı
- Complain: Müşteri son 2 yılda şikayet ettiyse 1, aksi takdirde 0
- MntWines: Son 2 yılda şaraba harcanan miktar
- MntFruits: Son 2 yılda meyvelere harcanan miktar
- MntMeatProducts: Son 2 yılda ete harcanan miktar
- MntFishProducts: Son 2 yılda balığa harcanan miktar
- MntSweetProducts: Son 2 yılda tatlılara harcanan miktar

- MntGoldProds: Son 2 yılda altına harcanan miktar
- NumDealsPurchases: İndirimle yapılan satın alma sayısı
- AcceptedCmp1: Müşteri 1. kampanyadaki teklifi kabul ederse 1, aksi takdirde 0
- AcceptedCmp2: Müşteri 2. kampanyadaki teklifi kabul ederse 1, aksi takdirde 0
- AcceptedCmp3: Müşteri 3. kampanyadaki teklifi kabul ederse 1, aksi takdirde 0
- AcceptedCmp4: Müşteri 4. kampanyadaki teklifi kabul ederse 1, aksi takdirde 0
- AcceptedCmp5: Müşteri 5. kampanyadaki teklifi kabul ederse 1, aksi takdirde 0
- Response: Müşteri son kampanyada teklifi kabul ederse 1, aksi takdirde 0
- NumWebPurchases: Şirketin web sitesi üzerinden yapılan satın alma sayısı
- NumCatalogPurchases: Bir katalog kullanılarak yapılan satın alma sayısı
- NumStorePurchases: Doğrudan mağazalarda yapılan satın alma sayısı
- NumWebVisitsMonth: Geçen ay şirketin web sitesine yapılan ziyaretlerin sayısı

Projede aşağıda sıralanmış kütüphaneler kullanılmıştır:

- Pandas (Veri işleme ve analiz)
- Numpy (Dizi ve matris işlemleri)
- Matplotlib (Görselleştirme)
- Seaborn (Görselleştirme)
- Plotly (3D Görselleştirme için kullanıldı.)
- Scipy (Sadece Dendrogram için kullanıldı.)
- Scikit Learn (Makine öğrenmesi)
- Sys ve Warnings (Modelden dönen uyarıları yakalamak için kullanıldı.)

2. Yöntem

2.1 Veri temizleme

Info fonksiyonu sayesinde ‘’Income’’ sütununda 24 adet eksik veri bulunduğu, ‘’Education’’, ‘’Martial Status’’ ve ‘’Dt_Customer’’ sütunlarının kategorik veriler içerdiği tespit edilmiştir. ‘’Income’’ sütunundaki 24 adet eksik veri için imputer tanımlamaya gerek duyulmadığından eksik verilerin bulunduğu satırlar doğrudan veri setinden çıkarılmıştır.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 29 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   2240 non-null   int64
 1   Year_Birth           2240 non-null   int64
 2   Education            2240 non-null   object
 3   Marital_Status       2240 non-null   object
 4   Income               2216 non-null   float64
 5   Kidhome              2240 non-null   int64
 6   Teenhome             2240 non-null   int64
 7   Dt_Customer          2240 non-null   object
 8   Recency              2240 non-null   int64
 9   MntWines             2240 non-null   int64
 10  MntFruits            2240 non-null   int64
 11  MntMeatProducts      2240 non-null   int64
 12  MntFishProducts      2240 non-null   int64
 13  MntSweetProducts     2240 non-null   int64
 14  MntGoldProds         2240 non-null   int64
 15  NumDealsPurchases    2240 non-null   int64
 16  NumWebPurchases      2240 non-null   int64
 17  NumCatalogPurchases  2240 non-null   int64
 18  NumStorePurchases    2240 non-null   int64
 19  NumWebVisitsMonth    2240 non-null   int64
 20  AcceptedCmp3         2240 non-null   int64
 21  AcceptedCmp4         2240 non-null   int64
 22  AcceptedCmp5         2240 non-null   int64
 23  AcceptedCmp1         2240 non-null   int64
 24  AcceptedCmp2         2240 non-null   int64
 25  Complain             2240 non-null   int64
 26  Z_CostContact        2240 non-null   int64
 27  Z_Revenue            2240 non-null   int64
 28  Response             2240 non-null   int64
dtypes: float64(1), int64(25), object(3)
memory usage: 507.6+ KB
```

Şekil 1. Info fonksiyonu

Şekil 2’de görüldüğü üzere ‘’Martial_Status’’ sütununda 8 farklı kategori, ‘’Education’’ sütununda ise 3 farklı kategori bulunmaktadır. Hali hazırda veri setinde fazla sütun bulunduğu için one hot encoder yöntemi kullanmak yerine ilgili sütunların etiketleri düzenlenerek label encoder yöntemine hazır hale getirilmiştir.

```
Martial Status kolonundaki toplam kategorik veriler
 Married     857
Together    573
Single      471
Divorced    232
Widow        76
Alone         3
Absurd        2
YOLO          2
Name: Marital_Status, dtype: int64

Education kolonundaki toplam kategorik veriler
 Graduation    1116
PhD             481
Master          365
2n Cycle        200
Basic            54
Name: Education, dtype: int64
```

Şekil 2. Kategorik sütunlardaki benzersiz değerler

İkiden fazla etikete sahip sütunlar çift etikete indirgenerek yeni etiketleri şu şekilde düzenlenmiştir.

- Martial Status:
    - Partner : Married, Together
    - Alone : Absurd, Widow, YOLO, Single
- Education
    - Undergraduate : Basic, 2n Cycle
    - Graduate : Graduatition, Master, PhD

Ardından "Martial_Status" sütununun değişen etiketleri ile duruma daha uygun olan "Living_With" sütunu oluşturulmuş olup binary'e dönüşen "Living_With" ve "Education" sütunları Scikit Learn kütüphanesinde bulunan Label Encoder yöntemi ile encode* edilmiştir.

"Dt_Customer'' sütunundaki müşterilerin işletmeye kayıt tarihleri kategorik vaziyettedir. Bu verileri datetime objesine çevirerek zaman damgası (timestamp)'na çevirmek yerine kayıt olan en yeni ve en eski müşteriler tespit edilerek buradan hareketle müşterilerin kayıtlı olduğu gün sayısı'nın tutulduğu ''Customer_For'' sütunu oluşturulmuştur.

2.1.2 Öznitelik Çıkarımı

Pandas ve Numpy kütüphaneleri kullanılarak;

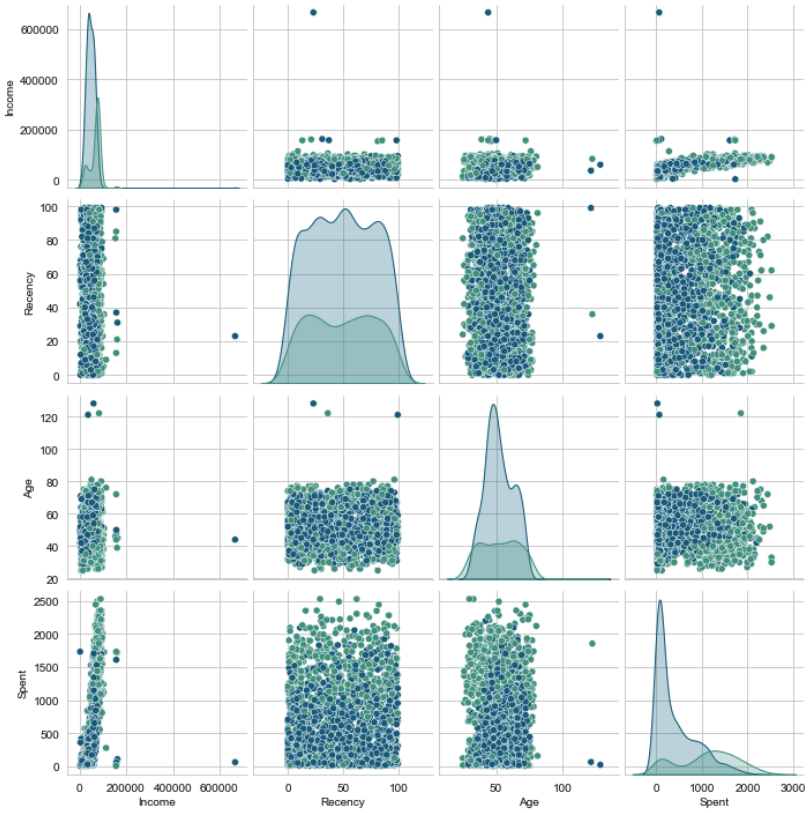
- Doğum yıllarının tutulduğu "Year_Birht" sütunundan hareketle "Age" sütunu oluşturulmuştur.
- "MntWines", "MntFruits", "MntMeatProducts", "MntFishProducts", "MntSweetProducts", "MntGoldProds" tek sütunda toplanmış olup "Spent" olarak isimlendirilmiştir.
- "Kidhome" ve "Teenhome" sütunlarından hareketle tek bir "Children" sütunu oluşturulmuştur.
- "Children" sütunundan elde edilen bilgiye göre ise ebeveynlik durumunu gösteren "Is_Parent" sütunu oluşturulmuştur.
- ''Living_With'' ve ''Children'' sütunlarından hareketle ''Family_Size'' sütunu oluşturulmuştur.

İşlenmiş, artık gerek kalmayan ve anlamsal olarak bağlantılı olmayan sütunlar; ''Martial_Status'', ''Dt_Customer'', ''Z_CostContact'', ''Z_Revenue'', ''Year_Birth'', ''ID'' veri setinden çıkarılmıştır.

2.2 Keşifsel veri analizi (Explority data analysis)

2.2.1 Pairplot

Şekil 3.'de çizdirilmiş pairplot incelendiğinde bazı verilerin marjinal dağılımlarının normale yakınsamadığı ve anomalik veriler içerme ihtimali olduğu görülmektedir.

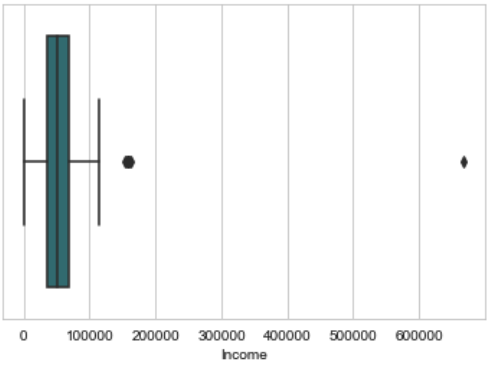


Şekil 3. (Pair Plot)

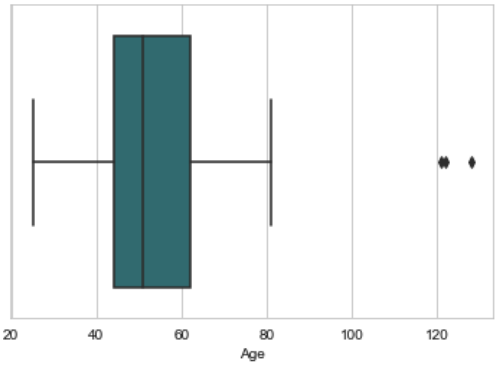
2.2.2 Anomali tespiti (Anomaly detection)

Anomali tespiti (aykırı değer tespiti) verilerin çoğunluğundan önemli ölçüde ayrışan, şüphe uyandıran verilerin tespit edilmesi işlemidir.

Bu durumu gözlemleyebilmek için şekil 4 ve şekil 5'de görüldüğü üzere boxplot yöntemiyle sapan değerler tespit edilmiştir.



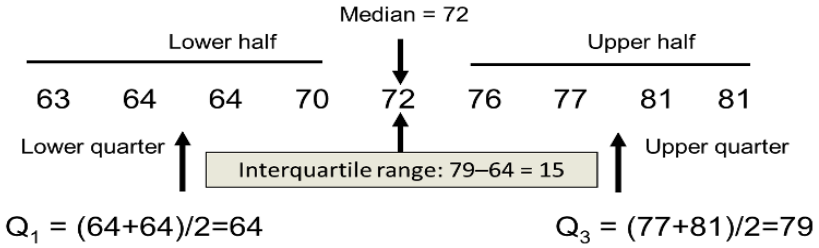
Şekil 4.



Şekil 5.

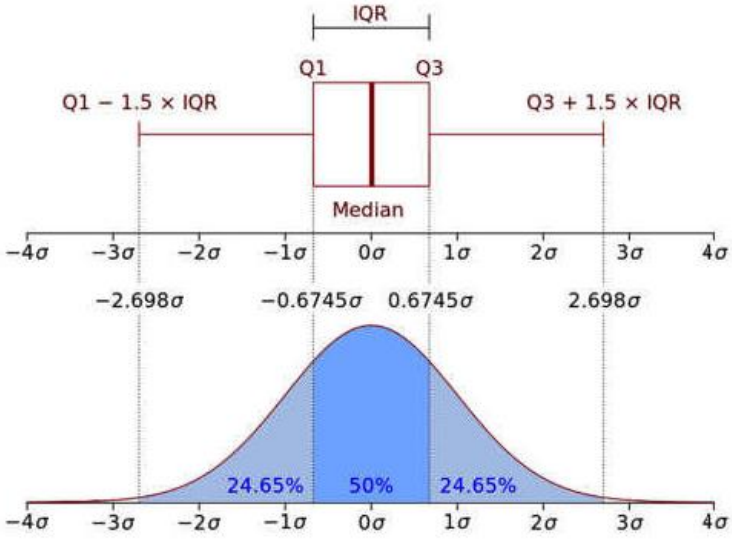
2.2.3 Anomalik verilerin temizlenmesi

Tespit edilen aykırı değerleri silmek için çeyrekler açıklığı (inter quantile range) yöntemi kullanılmıştır



Şekil 6.

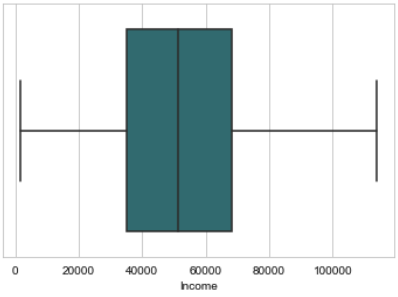
Şekil 6'da gösterildiği üzere çeyrekler açıklığı methodu ilgili sütunu küçükten büyüğe sıralayarak medyan noktasını belirler. (Q2) Medyan üzerinden alt ve üst yarımlara bölerek bu yarımlarında medyanları tespit edilir. Alt yarım'ın medyanı Q1(minimum ve ana medyan arasındaki ortalama değer), üst yarımın medyanı ise Q3 (maximum ve ana medyan arasındaki ortalama değer) olarak tanımlanır. Bu işlemden sonra IQR değeri; IQR=Q3 – Q1 hesaplanır.



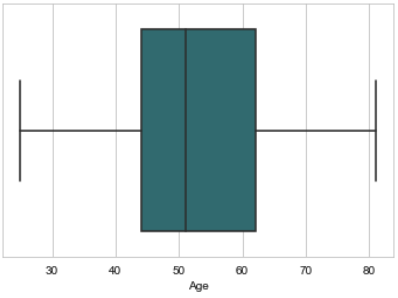
Şekil 7.

Aralığın alt ve üst sınırlarını tespit etmek üzere lowerbound ve upperbound değerleri şu şekilde hesaplanır; LB = Q1 – 1.5 x IQR UB = Q3 + 1.5 x IQR (Buradaki 1.5 sabiti veri setinin durumuna göre değişiklik gösterebilir.) Hesaplanan değerler sonucunda alt ve üst sınır dışında kalan değerler temizlenerek verideki anomalik durumlar ortadan kaldırılır. Şekil 7'de bu değerlendirmelerin dağılımda karşılık geldiği red noktaları görülmektedir.

Çeyrekler açıklığı methodu ile aykırı değerler ilgili sütunlardan çıkarıldıktan sonraki boxplot görünümü şekil 8 ve 9'da görülmektedir.



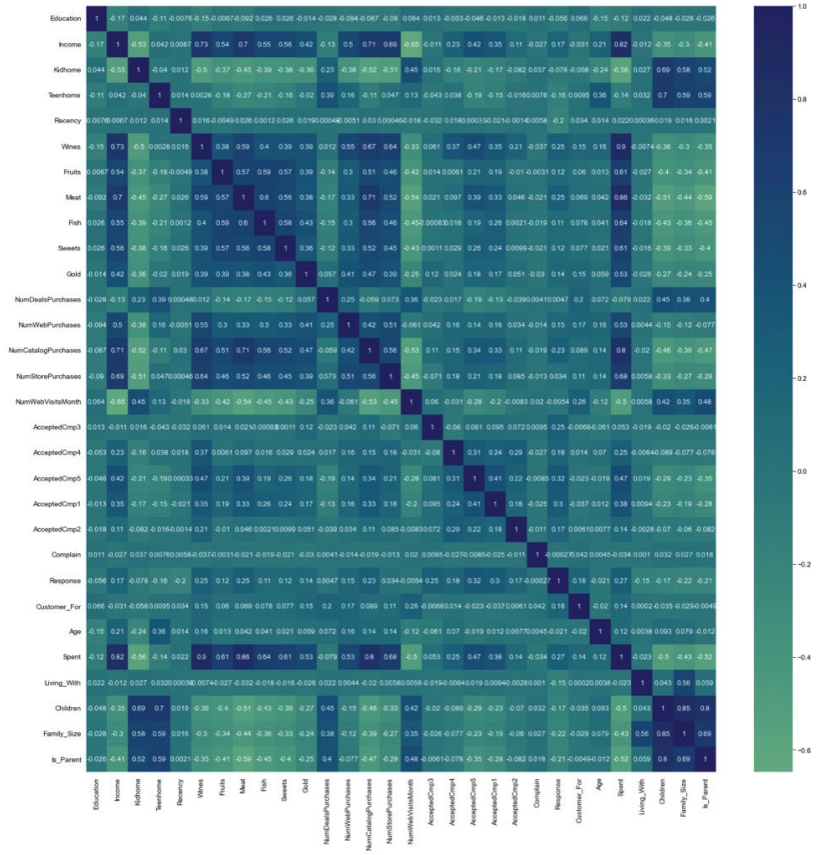
Şekil 8.



Şekil 9.

2.2.4 Korelasyon matrisi

Koyu tonlar pozitif korelasyonu temsil ederken, açık tonlar ise negatif korelasyonu temsil etmektedir. Detaylıca incelendiğinde öznitelik çıkarımı kısmında üretilen sütunlar ve bu sütunlara kaynak olan sütunlar ve mantıksal bağıntısı bulunan sütunlar (gelir ve şarap tüketimi gibi.) pozitif korelasyon göstermektedir. (Şekil 10)



Şekil 10. Korelasyon ısı haritası

2.3 Veri ön işleme

2.3.1 Sütunların düzenlenmesi

Veri ön işleme safhasında öncelikle veri setinin bir kopyası ‘’new_df’’ adlı değişkene alınmış olup bunun sebebi kümeleme işleminden sonra ortaya çıkan kümelerin ana veri setinde ilgili satırlarla ilişkilendirilerek görselleştirilecek olmasıdır. Bu işlemin ardından ‘’AcceptedCmp1’’, ‘’AcceptedCmp2’’, ‘’ AcceptedCmp3’’, ‘’ AcceptedCmp4’’, ‘’ AcceptedCmp5’’, ‘’ Complaim’’, ‘’ Response’’ sütunları yapılan testlerde (kümeleme denemeleri) anlamsal karmaşaya sebep olarak ortaya çıkan kümelerde mantıksal karmaşa ortaya çıktığından bu sütunlarda veri setinden çıkarılmıştır.

2.3.2 Veri ölçeklendirme

Veriler içerisindeki baskınlığı azaltarak işlem maliyetini azaltmak amacıyla ölçeklendirmeye ihtiyaç duyarız. Bu işlem için Standardizasyon gerçekleştirilmiş olup ortalama değerin 0, standart sapmanın ise 1 değerini alarak dağılımın normale yaklaştırıldığı bir metottur.

$$z = \frac{x_i - \mu}{\sigma}$$

Şekil 11.

Şekil 11'de de görüldüğü üzere ilgili veriden ortalama değer çıkarılarak varyans değerine bölünmesi ile elde edilir. Şekil 12'de veri seti'nin standardize edildikten sonra oluşan görüntüsü görülmektedir.

	Education	Income	Kidhome	Teenhome	Recency	Wines	Fruits	Meat	Fish	Sweets	...	NumCatalogPurchases
0	-0.359211	0.314651	-0.823405	-0.930767	0.310830	0.974566	1.548614	1.748400	2.449154	1.480301	...	2.628526
1	-0.359211	-0.254877	1.038757	0.906602	-0.380600	-0.874776	-0.638664	-0.731678	-0.652345	-0.635399	...	-0.588043
2	-0.359211	0.965354	-0.823405	-0.930767	-0.795458	0.355155	0.568110	-0.175957	1.336263	-0.149031	...	-0.230646
3	-0.359211	-1.206087	1.038757	-0.930767	-0.795458	-0.874776	-0.563241	-0.667380	-0.506392	-0.586763	...	-0.945440
4	-0.359211	0.322136	1.038757	-0.930767	1.555404	-0.394659	0.417263	-0.217292	0.150396	-0.003121	...	0.126750
...	...	...	...	...	...	...	...	...	...	...	...	...
2200	-0.359211	0.463624	-0.823405	0.906602	-0.104028	1.193879	0.417263	0.076644	0.077420	2.209853	...	0.126750
2201	-0.359211	0.598401	2.900920	0.906602	0.241687	0.295881	-0.663806	-0.621452	-0.688833	-0.659718	...	-0.230646
2202	-0.359211	0.258780	-0.823405	-0.930767	1.451690	1.783653	0.542969	0.237389	-0.105022	-0.367897	...	0.126750
2203	-0.359211	0.851004	-0.823405	0.906602	-1.417746	0.361082	0.090428	0.223611	0.770696	0.069834	...	0.841543
2204	-0.359211	0.060213	1.038757	0.906602	-0.311457	-0.658427	-0.588382	-0.479078	-0.652345	-0.635399	...	-0.588043

2205 rows × 23 columns

Şekil 12. (Ölçeklendirme işleminden sonra verilerin görünümü.)

2.3.3 Boyut indirgeme (Dimension Reduction)

Hali hazırda 23 adet sütunun bulunması verinin 23 boyutlu olduğu manasına gelmektedir. Bu boyut sayısı görselleştirmeyi imkansız hale getirirken işlem maliyetini oldukça arttırmaktadır. Bu olumsuz durumları ortadan kaldırmak için Temel Bileşen Analizi (Principal Component Analysis) yöntemi ile boyut indirgeme (dimension reduction) işlemi gerçekleştirilerek veri seti 3 boyuta indirgenmiş olup yeni veri seti şekil 13'de görüldüğü gibidir.

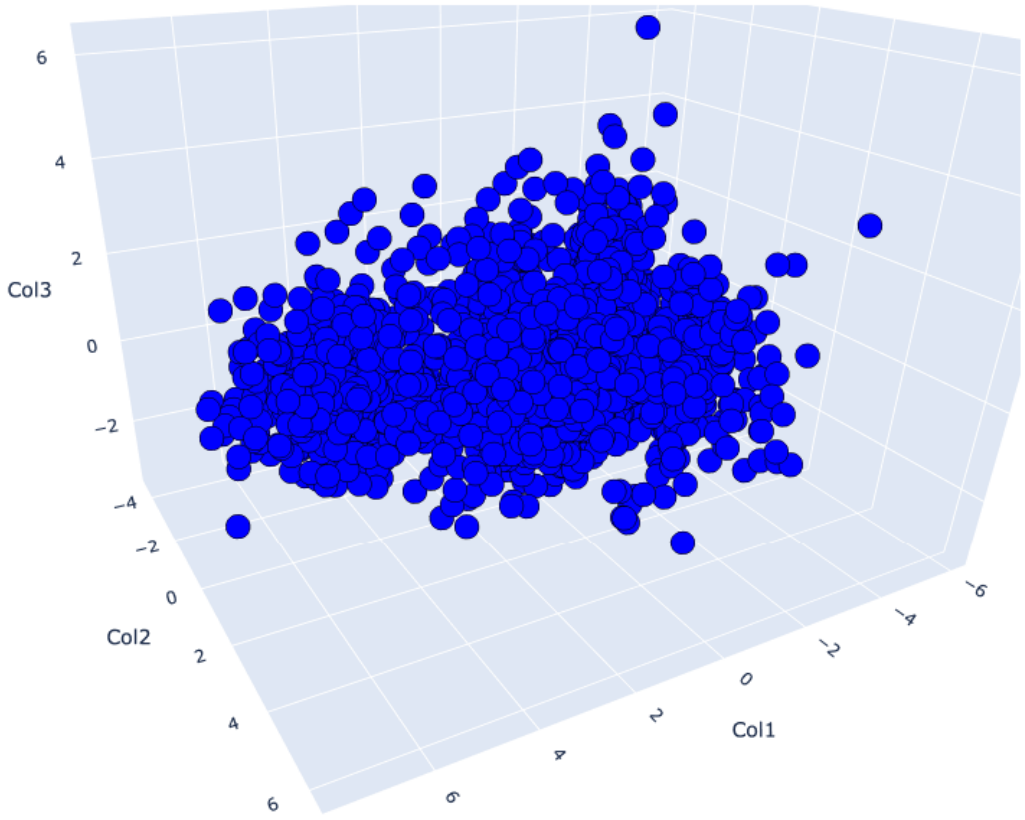
Temel bileşenler yaklaşımı bağımlılık yapısını yok etme ve boyut indirgeme amaçları için kullanılmaktadır. Tanıma, sınıflandırma, boyut indirgenmesi ve yorumlanmasını sağlayan, çok değişkenli bir istatistik yöntemidir. Bu yaklaşım verinin içindeki en güçlü örüntüyü bulmaya çalışır. Bu yüzden örüntü bulma tekniği olarak da kullanılabilir. Çoğunlukla verinin sahip olduğu çeşitlilik, tüm boyut takımından seçilen küçük bir boyut setiyle yakalanabilir. Verideki gürültüler, örüntülerden daha güçsüz olduklarından, boyut küçültme sonucunda bu gürültüler temizlenebilir.

	col1	col2	col3
0	5.024467	-0.209982	2.517811
1	-2.896486	0.047876	-2.105913
2	2.613544	-0.723597	-0.335315
3	-2.721422	-1.530335	-0.871340
4	-0.616060	0.292348	0.135577
...	...	...	...
2200	2.323871	2.367702	0.699915
2201	-3.047916	4.149750	-1.386849
2202	2.664088	-1.855911	0.192756
2203	1.566788	1.782320	-1.674938
2204	-2.707709	1.735478	-0.296739

2205 rows × 3 columns

Şekil 13. (Boyut indirgeme işleminden sonra veri setinin görünümü.)

3 boyuta indirgenmiş veriyi artık görselleştirmek mümkün olup verinin görünümü şekil 14'deki gibidir.

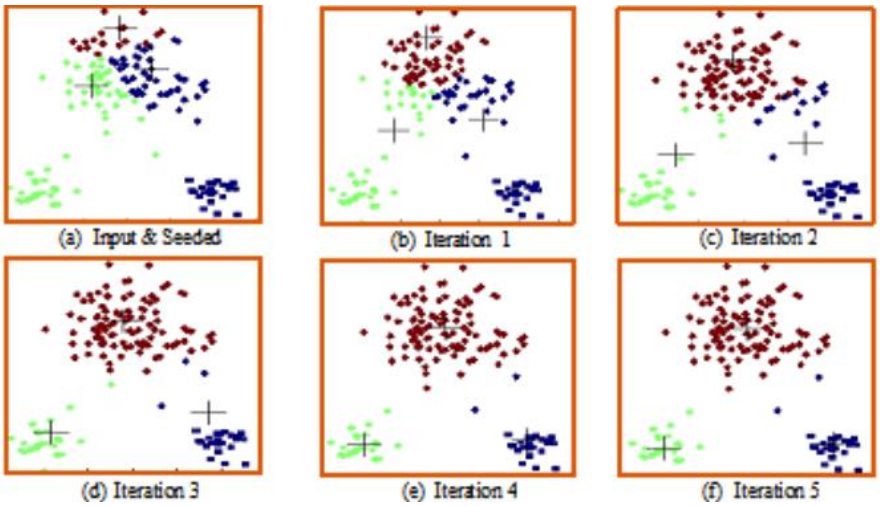


Şekil 14. (Plotly kütüphanesi ile 3 boyutlu görselleştirme.)

2.4 Modelleme

2.4.1 K-Means

K-Means kümeleme problemi, hesaplamalı geometri olarak tanımlanan mesafe bazlı kümeleme algoritmalarının en eski ve en önemlilerindendir. Amaç her nokta ile en yakın merkezi arasındaki toplam kare uzaklığı en aza indierecek merkezler belirlemektir. K-Means olarak anılan Lloyd algoritması, k adet rastgele merkez ile başlar. Her nokta daha sonra en yakın merkez ve her merkez, kendisine atanan tüm noktaların kütle merkezi olarak yeniden hesaplanır. Bu son iki adım süreç stabilize olana kadar şekil 15'te görüldüğü üzere tekrarlanır.



Şekil 15. K-means

K-Means algoritmasını çekici kılan doğruluğu değil, hızı ve basitliğidir. Aslında, K-Means'in rastgele olarak kötü kümeler ürettiği bir çok doğal örnek mevcuttur. Bu bağlamda model başarımını düşüren rastgele merkezlerin seçilmesi problemine çözüm olarak k-means++ algoritması başlatma methodu olarak kullanılması olası problemleri ortadan kaldırmaya yardımcı olacaktır.

K-means++ algoritması şu adımlarla özetlenebilir;
1- X'ten rastgeleseçilmiş bir c1 merkezi oluşturulur..
2- $D(x)^2/\sum x \in X D(x)^2$ Olasılıkla $x \in X$ seçerek yeni bir ci merkezi oluşturulur.
3- 2ci adım teararlanır
4- Ve standart K-means adımları ile devam edilir.

Kümeleme probleminde bir diğer problem optimal küme miktarının belirlenmesidir. Optimal küme sayısının belirlenmesinde farklı yaklaşımlar mevcuttur bunların arasında en popüler olanı ise Dirsek (Elbow) Methodudur.

Dirsek (Elbow) yaklaşımı, K-Meanstaki optimal küme sayısının bir fonksiyonu olarak gösterilen varyans yüzdelerine bakan bir tekniktir . Bu yaklaşım, çok sayıda kümenin seçilmesi gerektiği fikriyle ortaya çıkmış ve optimal küme sayısını bulmada yaygınca kullanılmaktadır..

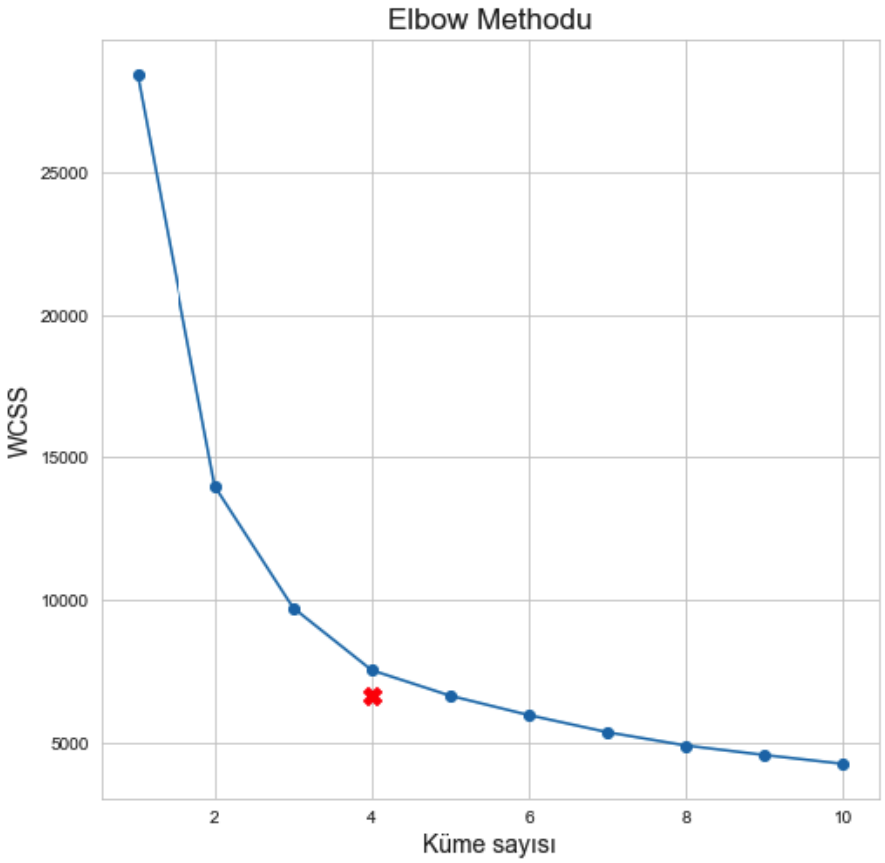
Kümeler tarafından gösterilen varyans yüzdesi, optimal küme sayısına karşı çizilir. İlk kümeler çok sayıda bilgi ekleyecektir, ancak belirli bir noktada, elde edilen marjinal küme sayısı önemli ölçüde düşecek ve grafikte bir açı sağlayacaktır. Bu noktada seçilen küme sayısı olan uygun "k" değerine Dirsek değeri denir. Elbow methodunun mantığı k = 2 ile başlamak, onu bir puan artırmaya devam etmek, kümeyi ve eğitimle birlikte gelen maliyeti hesaplamak olarak tanımlanabilir. 'k' için bir değerde, maliyet önemli ölçüde düşecektir ve bu noktadan sonra, k noktasını yükseltildiğinde maliyet yükselir. Maliyet düşüşünün bir maliyet artışına dönüştüğü noktada, dirsek olarak aradığınız k değeridir. Yani k = 4 noktasında bir dirsek vardır. Bu, optimal küme miktarının k = 4 olduğu anlamına gelir.

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$

Şekil 16. WCSS

Dirsek (Elbow) methodunun uygulanabilmesi için gerekli olan istatistiksel mesafe ölçütü grup içi kareler toplamı (Within Cluster Sum of Square) olarak tanımlanır; burada bir küme için tüm veri noktalarının (bireysel grupların veya grup ortalamalarının her birinin ortalamaları) arasındaki farkın kareleri wcss olarak adlandırılır. (Şekil 16)

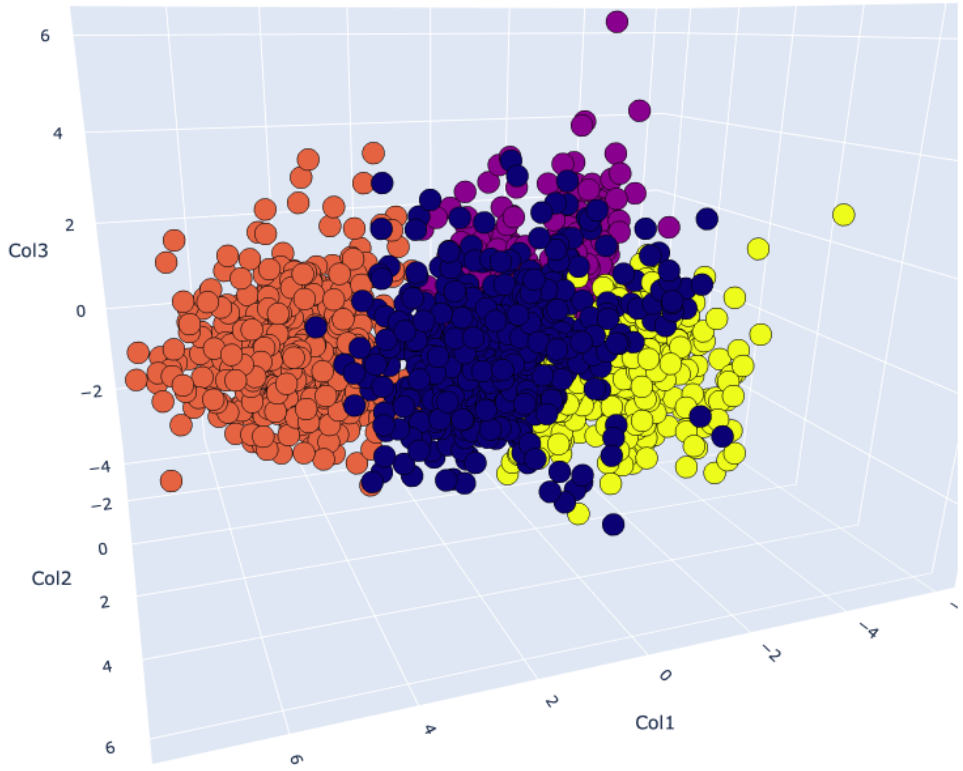
Hem K-Means hem de Dirsek yönteminin kombinasyonu ile optimal küme miktarı bulunabilir.



Şekil 17. (Dirsek Methodu)

Şekil 17'da görüldüğü üzere maliyet düşüşünün bir maliyet artışına dönüştüğü nokta olan 4 noktası optimal küme miktarını işaret etmektedir.

Şekil 18'de optimal küme miktarı olan 4 ve başlatma methodu olarak k-means++ kullanılarak veri seti 4 farklı kümeye bölünmüş ve Plotly kütüphanesi ile 3 boyutlu saçılım grafiği görülmektedir.



Şekil 18. (Kümeleme işleminin ardından verinin görüntüsü.)

2.4.2 Hiyerarşik Kümeleme (Hierarchical Clustering)

Hiyerarşik ve hiyerarşik olmayan kümeleme tekniklerinin her ikisinde de ortak amaç kümeler arasındaki farklılıkları ve kümeler içi benzerlikleri en yüksek düzeye çıkarmaktır. Yani küme içi homojenlik arttırılırken kümeler arası homojenlik ise azaltılmaktadır.

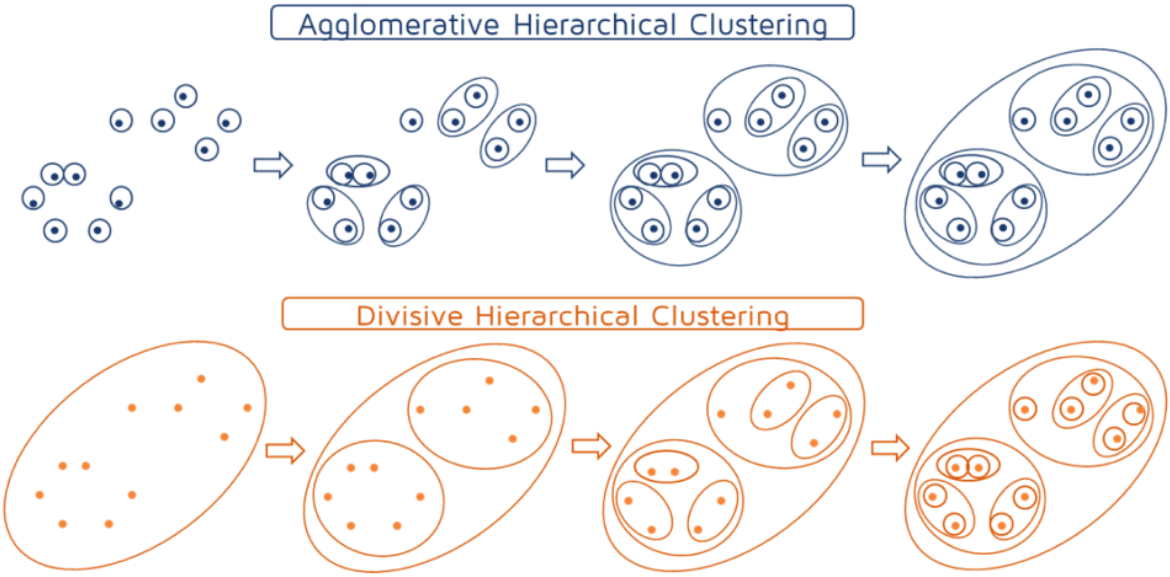
Hiyerarşik kümeleme yöntemi, kümelerden bir eleman silme ya da eklemeyle bir ağaca benzeyen yapı gösteren aşamalar grubudur (Ketchen and Shook, 1996: 444). Hiyerarşik kümeleme yöntemleri yöntemleri temel olarak birleştirici (agglomerative) ve ayrıştırıcı (divisive) olmak üzere iki ana başlıkta incelenir.

Birleştirici (Agglomerative) kümeleme her bir gözlemi bağımsız birer küme şeklinde ele alarak başlar. Ardından tekrarlı bir biçimde tüm gözlemleri kendine en yakın olan gözlem veyahutta veri kümesi ile bir küme oluşturmasını sağlar.

Ayrıştırıcı (Divisive) kümeleme yöntemlerinde başlangıçta tüm veri noktaları tek bir küme olarak ele alınır ve ardından yinelemeli olarak tüm gözlemler birbirinden bağımsız bir

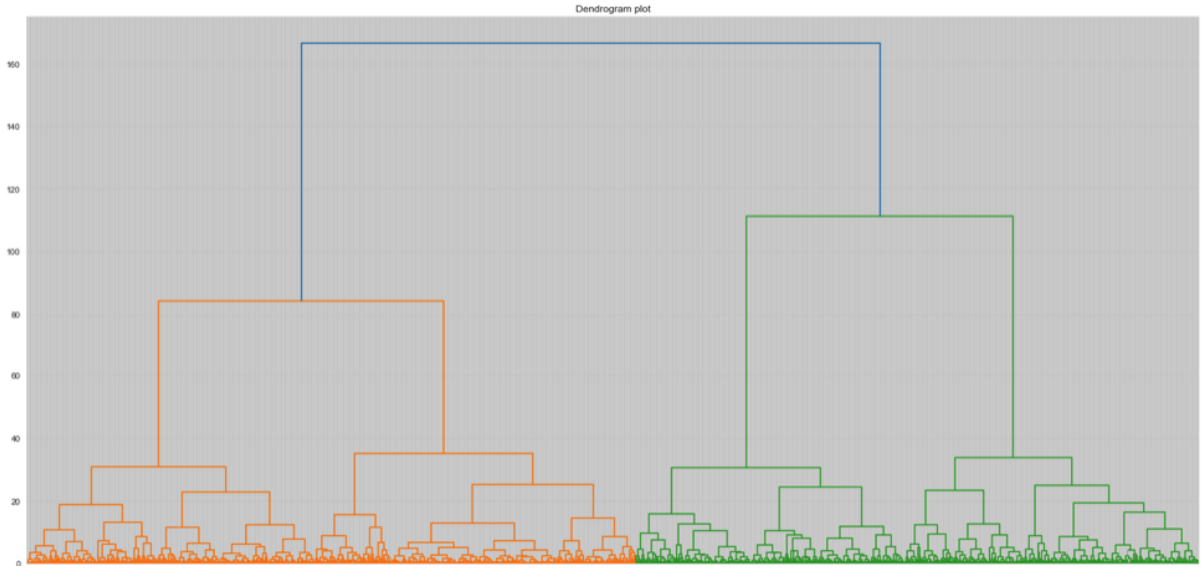
küme haline gelinceye kadar ayrıştırılır.

Bu iki yöntem arasındaki fark şekil 19'de açıkca görülmektedir.



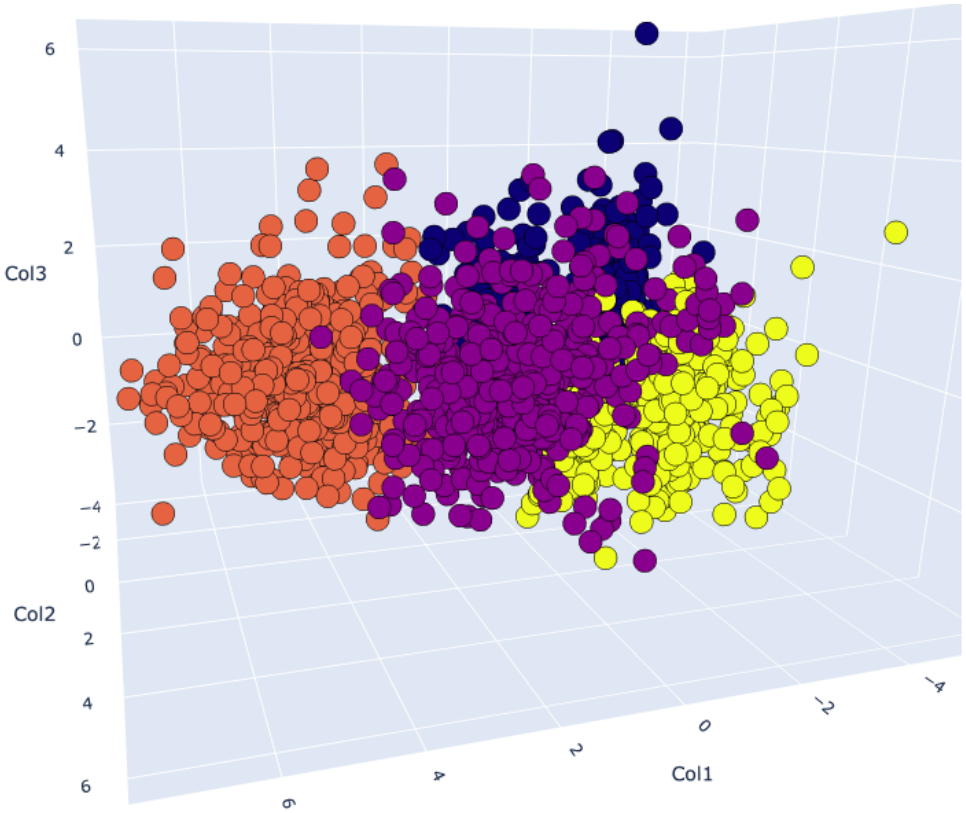
Şekil 19. Agglomerative ve Divisive kümeleme

Bu yukarı ve aşağı birleştirici veyahutta ayrıştırıcı hareketler dendrogramlar sayesinde görselleştirilerek kümeleme senaryoları gözlemlenir. Dendrogram üzerinde çizilecek bir yatay çizgi o çizgi üzerinde oluşacak kümeleme senaryosunu göstermektedir. (Şekil 20)

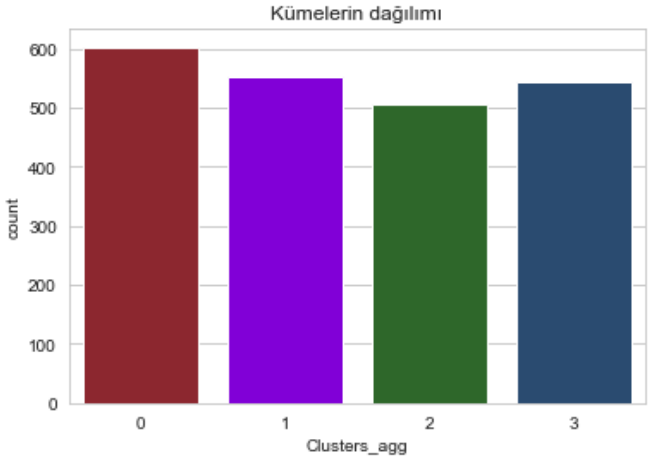


Şekil 20.(Dendrogram Plot)

Agglomerative kümeleme işleminden sonra kümelerin görüntüsü Plotly kütüphanesi ile 3 boyutlu görselleştirilmiştir. (Şekil 21.)



Şekil 21. (Agglomerative kümeleme işleminden sonra verinin görüntüsü.)



Şekil 22. (Kümelerdeki veri miktarı.)

Şekil 22'de görüldüğü üzere kümelerdeki gözlem miktarları Seaborn kütüphanesi yardımıyla görselleştirilmiştir. Kümelerdeki gözlem miktarlarının birbirine yakın olduğu grafik vasıtasıyla söylenebilir.

3.Bulgular

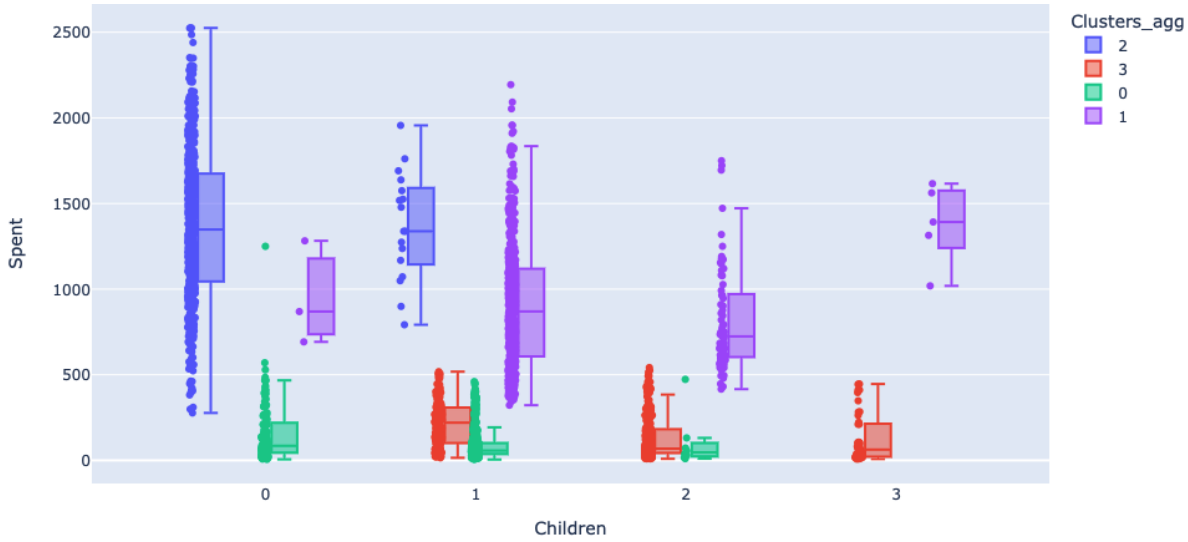
Farklı kümelere dahil bireylerin harcama ve gelir verileri kullanılarak Seaborn kütüphanesi ile saçılım grafiği (scatter plot) çizdirilerek grupların profilleri saptanmıştır. (Şekil 23.)

Şekil 23.

Grup profilleri:

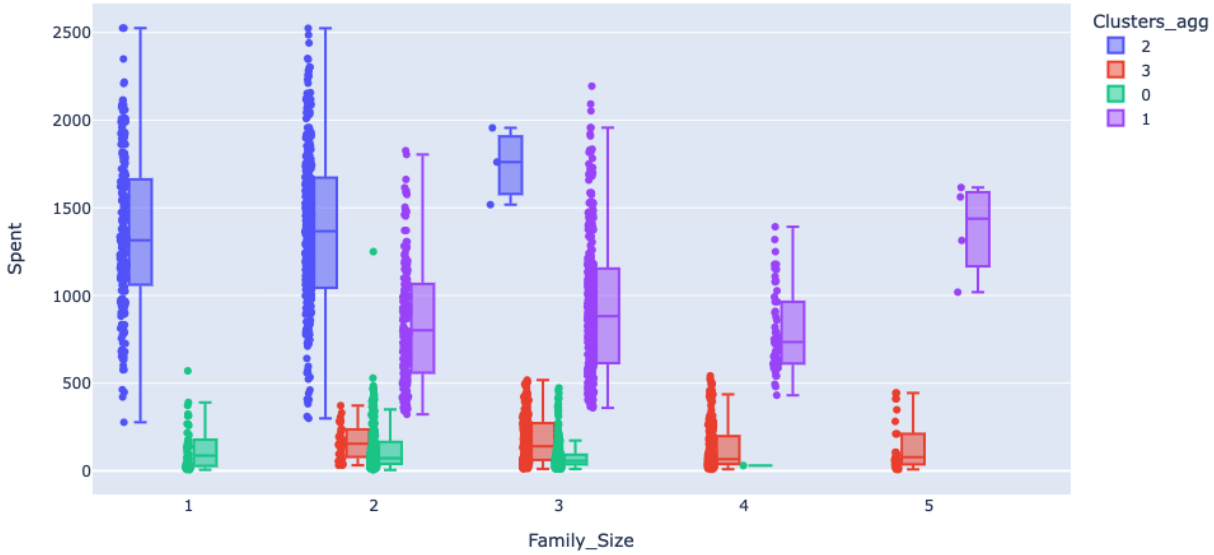
- **Grup 0:** Düşük gelir, düşük harcama
- **Grup 1:** Ortalama gelir, yüksek harcama
- **Grup 2:** Yüksek gelir, yüksek harcama
- **Grup 3:** Ortalama gelir, düşük harcama

Saptanmış olan harcama gruplarının sahip oldukları çocuk sayıları görselleştirilerek Grup 2: Yüksek gelir, yüksek harcama grubunun belirgin şekilde diğer gruplardan ayrışarak büyük çoğunluğunun çocuğun olmadığı ve geri kalan kısmının sadece 1 çocuğu olduğu saptanmıştır. (Şekil 24.)

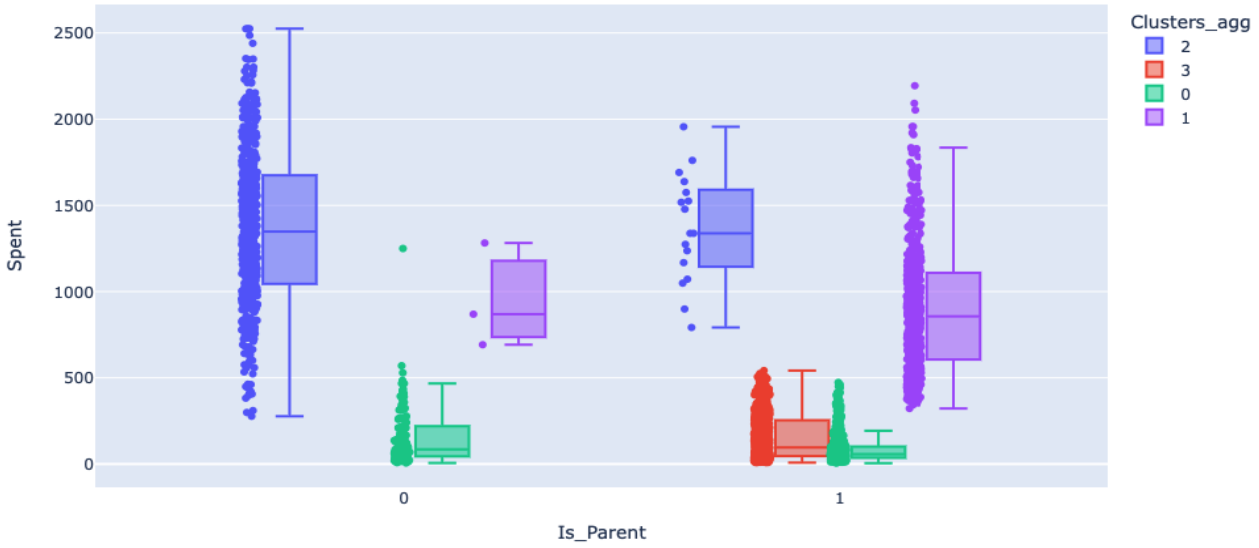


Şekil 24.

Harcama profillerinin aile üyeleri arasındaki farkı gözlemleyebilmek üzere ‘’Family_Size’’ ve ‘’Spent’’ sütunlarının kümeler arasındaki dağılımını gözlemlemek için Şekil 25’da görüldüğü üzere görselleştirilmiştir. Yüksek harcama profili olan Grup 2’nin diğer gruplardan ayrışarak ya yalnız yaşadıkları yada en fazla 1 partner ile yaşamayı tercih ettikleri görülmektedir. Ortalama gelir grubu olan grup 1’in baskın diğer grupların ise normal şekilde genellikle aile kurmayı tercih ettikleri görülmektedir.

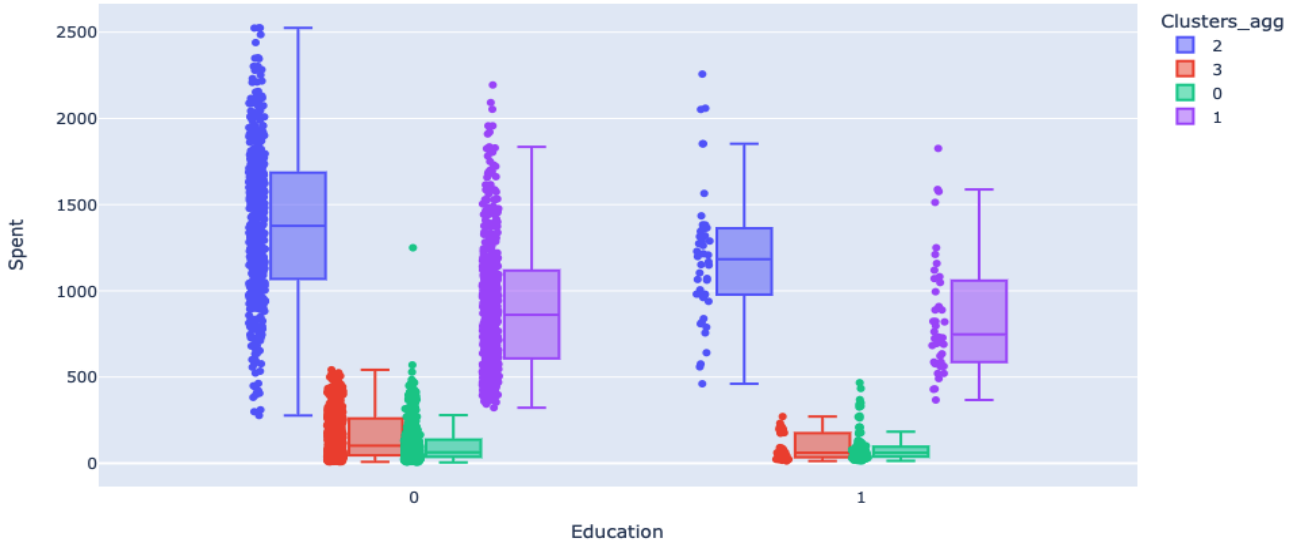


Şekil 25.



Şekil 26.

Veri ön işleme kısmında ‘’Is_Parent’’ sütununun encode edilmesinden ötürü 0 ile ebeveynlik durumunun olumsuz 1 ile ise olumlu olduğu gösterilmektedir. Grafikten hareketle yüksek gelir grubu olan Grup 2’nin diğer gruplardan belirgin şekilde ayrışarak ebeveynliği tercih etmediği şekil 26’de görülmektedir. Ortalama harcama profili olan grup 1 ‘in ebeveynliği tercih eden kısmının fazla olduğu görülmektedir.

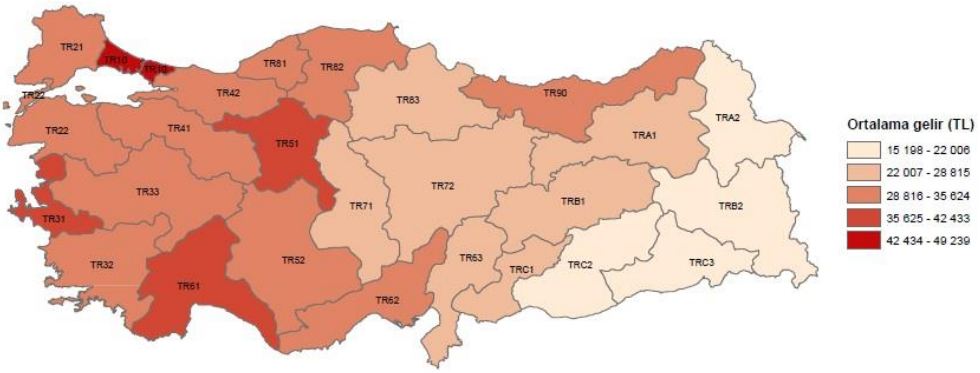


Şekil 27.

Veri ön işleme kısmında ‘’Education’’ sütununun encode edilmesinden ötürü kişinin 0 ile mezun olduğu, 1 ile ise mezun olmadığı şekilde 27’deki grafikte gösterilmektedir. Tüm grupların kendi içlerinde büyük çoğunluklarının mezun olan grupta yer aldığı gözlemlenmiştir.

4. Sonuç ve Tartışma

Gelir grupları incelendiğinde diğer gelir gruplarından yüksekliği ile ayrılan grubun bireysel yaşamaya diğer gruplardan daha çok önem verdiği bulgusu TÜİK'in ''Doğum İstatistikleri, 2020'' ve "Gelir ve Yaşam Koşulları Araştırması Bölgesel Sonuçları, 2020" raporları tarafından sunmuş olduğu Türkiye örnekleri ile de (projedeki veri seti Türkiye örnekleri içermemektedir.) desteklenmektedir. ''Yıllık ortalama eşdeğer hane halkı kullanılabilir fert geliri (TL), İBBS 2. Düzey, 2020'' (Şekil 28) ve '' İllere göre toplam doğurganlık hızı, 2020'' (Şekil 29) grafikleri incelendiğinde birbirlerinin tersi oldukları görülmektedir. Ortalama gelirin azaldığı noktalarda doğurganlık azalırken, ortalama gelirin arttığı noktalarda doğurganlık azalmaktadır.



Şekil 28. Yıllık ortalama eşdeğer hane halkı kullanılabilir fert geliri (TL), İBBS 2. Düzey, 2020 (TÜİK)

Şekil 29. İllere göre toplam doğurganlık hızı, 2020 (TÜİK)

Çalışma mevcut bulgular ile sonlandırılmış olup veri ön işleme safhasında birleştirilmiş; şarap, meyve, et, balık, tatlı, altın gibi farklı harcama kalemlerine ilişkin veriler ayrı ayrı incelenerek J. Duesenberry'nin ''Tüketim harcamalarına yön veren tüketimin sosyal anlamıdır.'' görüşünün gelir seviyesinin yükseldikçe marjinal tüketimin eğiliminin araştırılması açısından geliştirmeye açıktır.

5. Kaynakça

(1) Cömert, Z. (2015). Temel Bileşenler Analizine Genel Bir Bakış, Erişim tarihi: 05.01.2022, http://www.zafercomert.com/Medya/2015_05_21_2_152_9274a3ea.pdf#viewer.action=download

(2) Omar, T., Alzahrani A., Zohdy M. (2020). Clustering Approach for Analyzing the Student's Efficiency and Performance Based on Data, Journal of Data Analysis and Information Processing, Erişim tarihi: 05.01.2022, http://www.tubitak.gov.tr/sites/default/files/2204_proje_kitapcik.pdf

(3) Arthur D.(2006), k-means++: The Advantages of Careful Seeding, Technical Report. Stanford, Erişim tarihi: 05.01.2022,

http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf

(4) Ketchen, D.Jr. ve Shook, C.L. (1996). ''The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique'' Strategic Management Journal, 17(6), pp. 441-458.

(5) Yeşilbudak, M., Kahraman H. T., Karacan H. (2010), VERİ MADENCİLİĞİNDE NESNE YÖNELİMLİ BİRLEŞTİRİCİ HİYERARŞİK KÜMELEME MODELİ, Gazi Üniv. Müh. Mim. Fak. Der., Erişim tarihi: 05.01.2022,

https://dergipark.org.tr/tr/download/article-file/75925

(6) TÜİK (2020). Gelir ve Yaşam Koşulları Araştırması Bölgesel Sonuçları, 2020, Erişim tarihi: 09.01.2022

https://data.tuik.gov.tr/Bulten/Index?p=Gelir-ve-Yasam-Kosullari-Arastirmasi-Bolgesel-Sonuclari-2020-37405

(7) TÜİK (2020). Doğum İstatistikleri, 2020 Erişim tarihi: 09.01.2022

https://data.tuik.gov.tr/Bulten/Index?p=Dogum-Istatistikleri-2020-37229