

## DETERMINING CUSTOMER SEGMENTATION AND BEHAVIOR MODELS WITH DATABASE MARKETING AND MACHINE LEARNING

DOI: 10.17261/Pressacademia.2021.1409

JMML- V.8-ISS.2-2021(2)-p.89-111

**Orkun Berk Koca**

Beykent University, Graduate School, Beyoglu, Istanbul, Turkey.

[orkunkocaa@gmail.com](mailto:orkunkocaa@gmail.com), ORCID: 0000-0002-2862-1226

Date Received: February 20, 2020

Date Accepted: May 5, 2021

OPEN ACCESS



### To cite this document

Koca, B.O. (2021). Determining customer segmentation and behaviour models with database marketing and machine learning. Journal of Management, Marketing and Logistics (JMML), 8(2), 89 -111.

Permanent link to this document: <http://doi.org/10.17261/Pressacademia.2021.1409>

Copyright: Published by PressAcademia and limited licensed re-use rights only.

### ABSTRACT

**Purpose-** The study investigates the effect of data-based marketing, which is of great importance for today's businesses, on the creation of customer segments and on the development of marketing strategies for those segments. The big data study, which consists of real customer data analyzed in this direction, aims to identify consumer behaviors that are not similar enough to develop different strategies and to determine how the strategy development processes for customer segments can be done analytically.

**Methodology-** It was studied according to the 2018 order data, which is a data set of 24 million lines in total from an international pizza brand operating in the home delivery service field across the Turkey. In the study, K Means, Gaussian Mixture and DBSCAN algorithms are used for customer segmentation. The clustering and multiple regression analyzes were applied with the Phyton program.

**Findings-** In this study, in which the most used clustering algorithms in the literature were tested, due to the fact that the DBSCAN algorithm is not suitable for the data set used in the application, 91% of all data is assigned to a cluster in the cluster analysis and the remaining data are classified outlier. Accordingly, it has been observed that algorithms such as K Means or Gaussian Mixture give better results in studies where there is no demographic data and behavioral characteristics form the main mass of the data. In addition, clusters exhibiting similar behavior were identified in multiple regression analyzes, in which the forward-looking behaviors of the formed clusters were analyzed, and valuable clusters in the sub-clusters were discovered.

**Conclusion-** While this study covers the steps of big data and data mining, it also covers all end to end processes with multiple regression analyzes to create customer clusters and determine the future behavior models of the determined clusters. In this direction, it offers practitioners an exemplary model and strategy determination methodology.

**Keywords:** Database marketing, machine learning, customer clusters, segmentation, k-means

**JEL Codes:** M30, M31, C38

## VERİ TABANLI PAZARLAMA VE MAKİNE ÖĞRENMESİ İLE MÜŞTERİ BÖLÜMLEME VE DAVRANIŞ MODELLERİNİN BELİRLENMESİ

### ÖZET

**Amaç -** Çalışma, günümüz işletmeleri açısından önemi son derece artan veri tabanlı pazarlamanın, müşteri bölümleri oluşturulmasına ve oluşan bölümlere yönelik pazarlama stratejilerinin geliştirilmesine etkisini araştırmaktadır. Bu doğrultuda analiz edilen gerçek müşteri verilerinden oluşan büyük veri çalışması, birbirinden farklı stratejiler geliştirebilecek kadar benzer olmayan tüketici davranışlarını belirlemesi ve müşteri bölümlerine yönelik strateji geliştirme süreçlerinin nasıl analitik olarak yapılabileceğini tespit etme amacını taşımaktadır.

**Yöntem -** Türkiye genelinde evlere servis alanında faaliyet gösteren uluslararası bir pizza markasının 2018 yılına ait sipariş verileri baz alınarak, toplam 24 milyon satır uzunluğunda veri seti ile çalışılmıştır. Çalışmada müşteri bölümlendirmesi için K Means, Gaussian Mixture ve DBSCAN algoritmaları kullanılmıştır. Söz konusu kümeleme ve çoklu regresyon analizleri Phyton programı ile uygulanmıştır.

**Bulgular -** Literatürde en çok kullanılan kümeleme algoritmalarının test edildiği bu çalışmada, DBSCAN algoritmasının, uygulamada kullanılan veri setine uygun olmaması nedeniyle tüm verinin %91'ini bir kümeye atarak geri kalan verileri, outlier farklı bir ifade ile aykırı olarak sınıflandırmıştır. Bu doğrultuda içerisinde demografik verinin bulunmadığı, davranışsal özelliklerinin verinin anakütlesini oluşturduğu çalışmalarda K Means veya Gaussian Mixture gibi algoritmaların daha iyi sonuç verdiği gözlenmiştir. Bununla birlikte oluşturulan kümelerin ileriye yönelik davranışlarının analiz edildiği çoklu regresyon analizlerinde, benzer davranış sergileyen kümelerin tespiti ile alt kümelerde bulunan değerli kümelerin keşfi sağlanmıştır.

**Sonuç -** Bu çalışma, büyük veri ve veri madenciliği adımlarını kapsarken, bununla birlikte müşteri kümelerinin oluşturulması, belirlenen kümelerin ileriye yönelik davranış modellerinin belirlenmesi için yapılan çoklu regresyon analizleri ile uçtan uca tüm süreçleri kapsamaktadır. Bu doğrultuda uygulayıcılara örnek bir model ve strateji belirleme metodolojisi sunmaktadır.

**Anahtar Kelimeler:** Veri tabanlı pazarlama, makine öğrenmesi, müşteri kümeleri, bölümlendirme, k-ortalamalar

**JEL Kodları:** M30, M31, C38

## 1. GİRİŞ

İşletmeler bilginin ulaşılabilir olması, maliyetinin azalması ve dolayısıyla değerinin artması ve müşteri odaklı pazarlama anlayışının gelişmesiyle birlikte, ilişkide buldukları müşteriler hakkındaki çeşitli verileri toplamak, depolamak ve bu verileri yararlı bir şekilde bilgiye dönüştürerek kullanmak amacıyla birtakım çalışmalara yönelmişlerdir. Bu çalışmaları inceleyen veri tabanlı pazarlama sayesinde işletmeler, hedef müşterilerine ait verilerden yola çıkarak etkili stratejiler geliştirmektedirler.

Günümüzde bilgisayar teknolojisindeki hızlı değişim ve ilerlemelerle birlikte yüksek miktarda verinin güvenilir bir şekilde depolanma imkanı elde edilmiştir. Dünya üzerinde katlanarak artan bu verinin, fayda sağlayan, anlamlı bir bilgiye dönüşmesi için artık disiplinler arası çalışmalar hızla artmaktadır. Pazarlama alanında ise bu yeni bilgi çağının gerekliliği olarak bilgi teknolojileri ile yakından çalıştığı gözlenmektedir. Makine öğrenmesi algoritmaları, müşteri davranışlarından yola çıkarak müşterilerin kümelerine ayrılması, öneri programlarının oluşturulması, yeni ürün geliştirilmesi, fiyatlama ve kayıp müşteri analizi gibi birçok konuda yeni stratejilerin oluşturulmasına katkı sağlamaktadır.

Makine öğrenmesinin pazarlama alanında son dönemde kullanımı hızla artmaya başlamıştır. Özellikle pazarlamanın en yaygın uygulama alanı olan perakende sektöründe perakendecilerin topladıkları veri, günümüzde büyük boyutlarla ifade edilmektedir. Üreticilerin akıllı sistemlerle stok uygulamaları ve lojistik çözümleri gibi faaliyetlerle perakendeci ile birebir ilişki kurması da verinin boyutlarını daha çok arttırmaktadır. Rekabete ayak uydurmak ve müşterilerine daha iyi hizmet sunmak amacı ile hareket eden perakende sektörü, POS (Point of Sales) sistemleri dışında, akıllı sensörler, kablosuz ağlar, vb. sistemler ile veri toplayarak kendi bulut sistemlerini oluşturmaya başlamıştır. Ayrıca firmalar, sosyal medya ve tüketici bloglarında bulunan tüketici yorumları, firma web sitelerini ziyaret verileri, müşterilerin konum tabanlı kullanıcı bilgilerine ilişkin dış kaynaklı verileri de elde edebilmekte, müşterilerini sürekli gözetim altında tutmakta ve bu verilerden faydalı çıkarsamalarda bulunmak amacıyla çeşitli analizler yapmaktadır.

## 2. LİTERATÜR TARAMASI

Veri tabanlı pazarlama, tüketicilere ait yaş, meslek, cinsiyet, eğitim durumu, kültürel, ekonomik vb. verileriyle, alışveriş eğilimlerine yönelik verilerin, bilgi teknolojileri aracılığı ile izlenmesi, işlenmesi ve çözümlenmesi sonucunda geliştirilen pazarlama faaliyetleridir (Verhoef, Spring, Hoekstra, Leeflang, 2003). Kotler ve Armstrong'a (1999) göre müşteri veri tabanları; "organize edilmiş, kapsamlı, içinde coğrafik, demografik, psikolojik ve davranışsal bilgilerin bulunduğu verilerdir. Veri tabanlı pazarlama ise, müşteri veri tabanlarının ya da diğer veri tabanlarının kullanılmasına dayanan pazarlama uygulamalarıdır". Müşteriler ile ilgili verilerin toplandığı müşteri veri tabanları, bu verilerin çalıştırıldığı ve bilgiye dönüştürüldüğü yer olarak tarif edilebilmektedir. Esas olarak bilgiye ve uzun dönemli müşteri ilişkilerine yoğunlaşan veri tabanlı pazarlama, tüm bu faaliyetleri gerçekleştirmek üzere oluşturulan pazarlama çabaları olarak belirtilmektedir. (Jackson, Wang, 1996).

Veri tabanlı pazarlamanın farklı bir tanımına göre, müşteri odaklı, mevcut ve potansiyel müşteriler için uzun vadeli bire bir pazarlama stratejilerinin uygulanmasına imkan veren, sadık müşteriler yaratan ve sağlayan pazarlamadaki, bilgi teknolojisi uygulaması olarak özetlenmektedir (Gülcan, 2000). Başka bir anlatım ile veri tabanlı pazarlama, arzulanan hedef müşteriler ile ilgili devamlı veri toplanması ve sınıflandırılması yoluyla öğrenilen yeni bilgilerin müşterilere yönelik iletişimlerde, uzun vadeli müşteri sadakati yaratması amacıyla kullanılmasıdır.

Veri tabanlı pazarlama, müşterilerin geçmişteki ve şu andaki bilgilerinin araştırılması yolu ile müşterilerle uzun süreli ilişkiler kurulması için oluşturulan pazarlama stratejilerinin geliştirilmesine ve uygulanmasına yardımcı olan sistemdir (Hoda ve Jocumens 2003). Haşiloğlu ve arkadaşlarına göre, "veri tabanlı pazarlamayı uygulamak için verilerin, formasyona, bir sonraki aşamada da yararlı bilgiye dönüştürülmesi gerekmektedir. Böylelikle, verileri stratejik olarak değerlendirerek potansiyel müşterilerin tespit edilmesine, müşteri memnuniyetinin sağlanmasına ve sürekli müşteri olmalarına destek olur" (Haşiloğlu, Sezgin, Bardakçı, 2008). Veri tabanlı pazarlama genel olarak üç amaçla kullanılmaktadır. Birinci olarak potansiyel müşteri ile ilgili ve pazarlama faaliyetleri için ihtiyaç duyulan verileri elde etmektir. İkinci olarak, verilerinden çıkartılan bilgilerle, farklı müşteri kümeleri için stratejiler oluşturmaktır. Üçüncü ise müşteri bağlılığı oluşturarak, satın alma faaliyetlerinin devamlı hale gelmesini sağlamaktır (Haşiloğlu, Sezgin, Bardakçı, 2008).

Veri tabanlı pazarlamanın pazarlama stratejileri geliştirme ve uygulama imkanı sağlaması, veri tabanlı pazarlama kullanan işletmelere kolaylıklar sağlamaktadır. Veri tabanlı pazarlamayı kullanan işletmeler amaçları, politikaları ve gördükleri işlevler doğrultusunda kendilerine özgü bir yol seçerler. Bu nedenle veri tabanlı pazarlamanın tüm özelliklerini, onu kullanan işletmelerin hepsinde görmek mümkün değildir. Ancak genel olarak potansiyel müşteri harcamalarının tespiti (potential), çapraz satışların değerlendirilmesi (cross sell), yaşam boyu değer belirlenmesi (lifetime value), cüzdan paylaşımı (share of wallet), müşteri alışveriş sıklığının korunması (attrition) alanlarında stratejiler geliştirmek üzere kullanılmaktadır (Koslowsky, 1999).

Makine öğrenmesi ve yapay zeka terimi ilk olarak 1950 yılında Turing'in (1950) "Computing Machinery and Intelligence" adlı makalesinde, makineler düşünebilir mi sorusunu ortaya atması ile çıkmıştır. Birçok araştırmacı ve akademisyen bu sorunun çevresinde makinelere farklı öğrenimler kazandırabilecek alanları geliştirmiştir. Alpaydın'a göre (2013): "Günümüzde bilgisayar

teknolojisindeki hızlı değişim ve ilerlemelerle birlikte yüksek miktarda verinin güvenilir bir şekilde depolanma imkanı elde edilmiştir. Dünya üzerinde üretilen veri katlanarak artmakta ve gözlemlenen verinin altında yatan bir süreç bulunmaktadır. Bu sürecin tamamen ortaya çıkarılması mümkün olamamakla birlikte, süreci anlamaya yaklaşacak bir model inşası hedeflenmektedir. Söz konusu sürece dair yakınsamalar, veri içerisindeki modelleri ve düzenleri anlayabilmek için önem arz etmektedir. Verinin akış düzenini sağlayan yapıya ait bu model ve düzenlere dair edinilen fikirler, yakınsamalar geleceğe dair öngöründe bulunulmasına imkan sağlamaktadır. Ayrıca, gelecek için tutarlı tahminler yapabileceği gibi geçmişe yönelik anlamlı sonuçlar da çıkartabilmektedir”.

Makine öğrenmesi, örnek veri seti ya da geçmiş veriler üzerinden bir performans kriterinin optimize edilmesi için bilgisayarları programlamaktadır (Alpaydın, 2013). Başka bir deyişle, verinin anlaşılabilir aksiyonlara dönüştürülmesini sağlayan, bilgisayar algoritmalarının geliştirilmesine dayalı çalışma alanı olarak tanımlanmaktadır (Cui, Wong, Lui, 2006). Söz konusu çalışma alanı; ulaşılabılır veri, istatistiksel metodlar ve gelişen hızlı sınıflama gücü olmak üzere bu üç temel sac ayağı üzerinde yükselmektedir (Lantz, 2013). Söz konusu üç temel dayanak, birbirinin gelişimine bağlı olarak beraber büyümektedir.

Literatürde pek çok makine öğrenmesi algoritması bulunmaktadır. Bunlardan bazıları; basit (naive) bayes sınıflandırıcı, karar ağaçları, yapay sinir ağları, k-en yakın komşu algoritması, lojistik regresyon analizi, k-ortalamlar (k-means) algoritmasıdır. Bu algoritmaların bir kısmı kümeleme ve sınıflandırma yaparken bir kısmı ise ileriye yönelik tahminlerde bulunmak için kullanılmaktadır. Makine öğrenmesi çeşitleri, eğitim verilerinin türleri, eğitim verilerinin elde edilmesi ve öğrenme algoritmasını değerlemek için kullanılan test verileri bakımından farklılık gösterir (Mohri, Rostamizadeh, Talwalkar, 2012). Makine öğrenmesi algoritmaları öğrenim özellikleri bakımından üçe ayrılır; gözetimli öğrenme, gözetimsiz öğrenme ve pekiştirmeli öğrenme (Canepa, 2016). Bu üç öğrenim literatürde sıklıkla kullanılmakla birlikte diğer öğrenme yöntemleri ise yarı gözetimli öğrenme, transdüktif akıl yürütme, çevrimiçi öğrenme ve aktif öğrenmedir (Mohri, Rostamizadeh, Talwalkar, 2012).

Kümeleme, bir gözetimsiz öğrenme (unsupervised learning) yöntemidir (Yılmaz, Patır, 2011). Amaç, elemanların birbirlerine çok benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere (gruplara) bölünmesidir (Durmuş, İplikçi, 2007). Sınıflamaları hakkında açık bilgi bulunmayan durumlarda, topluluğa ilişkin tahminlerin yapılmasında yararlanılan bir yöntemler grubu olarak tanımlanan kümeleme analizi, araştırmacıya, üzerinde çalışılan herhangi bir veri setindeki benzer (homojen) birey gruplarını bulma, kendi içinde türdeş fakat diğerlerinden farklı olacak biçimde kümelere ayırma olanağı tanıyan birçok değişkenli istatistiksel analiz tekniğidir. Segment analizi ve taksonomi analizi olarak da adlandırılan kümeleme analizinin genel amacı, gruplanmamış verileri benzerliklerine göre sınıflandırmak ve araştırmacıya uygun, işe yarar ve özetleyici bilgiler elde etmede yardımcı olmaktır (Harrigan, 1985; Sambamoorthi, 1999; Ketchen, Shook, 1996).

Makine öğrenmesinin pazarlama alanında son dönemde kullanımı hızla artmaya başlamıştır. Özellikle pazarlamanın en yaygın uygulama alanı olan perakende sektöründe perakendecilerin topladıkları veri, günümüzde büyük boyutlarla ifade edilmektedir. Üreticilerin akıllı sistemlerle stok uygulamaları ve lojistik çözümleri gibi faaliyetlerle perakendeci ile birebir ilişki kurması da verinin boyutlarını daha çok arttırmaktadır. Rekabete ayak uydurmak ve müşterilerine daha iyi hizmet sunmak amacı ile hareket eden perakende sektörü, POS (Point of Sales) sistemleri dışında, akıllı sensörler, kablosuz ağlar, vb. sistemler ile veri toplayarak kendi bulut sistemlerini oluşturmaya başlamıştır. Ayrıca firmalar, sosyal medya ve tüketici bloglarında bulunan tüketici yorumları, firma web sitelerini ziyaret verileri, müşterilerin konum tabanlı kullanıcı bilgilerine ilişkin dış kaynaklı verileri de elde edebilmekte, müşterilerini sürekli gözetim altında tutmakta ve bu verilerden faydalı çıkarsamalarda bulunmak amacıyla çeşitli analizler yapmaktadır.

### 3. VERİ VE METODOLOJİ

Uygulama, günümüz işletmeleri açısından önemi son derece artan veri tabanlı pazarlamanın, müşteri bölümleri oluşturulmasına ve oluşan bölümlere yönelik pazarlama stratejilerinin geliştirilmesine etkisini araştırmaktadır. Literatürdeki araştırmalar incelendiğinde veri tabanlı pazarlamanın müşteri bölümlere ve pazarlama stratejilerinin geliştirilmesindeki önemi vurgulanmaktadır (Verhoef, Spring, Hoekstra, Leeflang, 2003; Dibb, Meadows, 2004; Chaffey, 2009). Ayrıca müşteri tatmini ve müşteri bağlılığını yükselterek sadık müşteriler yarattığı ve farklı müşteri bölümleri için stratejiler geliştirilmesine imkan tanıdığı tespit edilmiştir (Gülcan, 2000; Haşiloğlu, Sezgin ve Bardakçı, 2008).

Bu çerçevede uygulamada şu temel sorunsala odaklanılmıştır: Büyük veriden yararlanarak pazarlama bölümleri nasıl daha etkin ve daha hassas şekilde bölümlenebilir? Literatürde yer alan pazar bölümlere kriterleri nasıl kullanılabilir? Tüketim davranışları birbirinden farklı olan kendi içinde nispeten homojen pazar bölümleri için pazarlama stratejileri ayrı ayrı nasıl geliştirilebilir?

Türkiye genelinde evlere servis alanında faaliyet gösteren uluslararası bir pizza markasının 2018 yılına ait sipariş verileri baz alınarak, hangi ürün ve ürün gruplarının, ne sıklıkta, hangi saat aralıklarında ve hangi sipariş kanalından vb. sorulara yanıt bulacak şekilde araştırılmasıyla müşteri kümeleri ve davranış modelleri tespit edilmiştir. Toplam 24 milyon satır uzunluğunda veri seti ile çalışılmıştır. Bu veri setindeki özellikler çeşitli işlemlerden geçtikten sonra 52 adet özellik ile başlanarak analizler yapılmıştır.

Bu doğrultuda literatürde en çok tercih edilen üç farklı kümeleme algoritması test edilmiştir. Bunlar; K-Means, Gaussian Mixture ve DBSCAN algoritmalarıdır. K means kümeleme yöntemi literatürde katı c ortalamalar olarak da bilinen bir yöntemdir. Bu açıdan yöntem tekrar ifade edilecek olursa; bu yöntem n adet vektör grubunu c adet gruba ayırmaya çalışan bir yöntemdir. MacQuenn (1967), en yakın değerlere sahip her elemanı, kümelere ayırabilecek algoritmayı tanımlamak için k means (ortalamalar) terimini ortaya atmıştır. K-means algoritmasında uzaklık ölçütü olarak genellikle öklit uzaklığı kullanılırken, algoritmanın değerlendirilmesinde en yaygın olarak, karesel hata kriteri SSE (sum of squared error) kullanılmaktadır. En düşük SSE değerine sahip kümeleme sonucu, en iyi sonucu verir. Nesnelerin buldukları kümenin merkez noktalarına olan uzaklıklarının karelerinin toplamıdır.

$$\mu_i = \frac{1}{|S_j|} \sum_{x_i \in S_j} x_i \quad \arg \min_s \sum_{j=1}^K \sum_{x_i \in S_j} \|x_i - \mu_j\|^2$$

Bu kriterler sonucu k tane kümenin olabildiğince yoğun ve birbirinden ayrı sonuçlanması hedeflenmektedir. Algoritma, karesel hata fonksiyonunu azaltacak k parçayı belirlemektedir (Kantardzic, 2011). Bu kümeleme algoritmalarından yola çıkarak anlamlı kümelerin tespiti için ise Sillhouete, Davies Bouldin ve Calinski Harabasz indeksleri test edilmiştir. Sillhoutte, veri kümeleri içerisindeki tutarlılığın yorumlanması ve doğrulanması yöntemidir. Yöntem her bir nesnenin ne kadar iyi bölümlendirildiğine dair kısa ve açık grafik gösterimi sağlar. Sillhouete katsayısı, bir nesnenin diğer kümelere kıyasla kendi kümesine ne kadar benzediğinin bir ölçütüdür (Rousseeuw, 1987). Bu değer -1 ile +1 arasındadır. Değer 1'e ne kadar yakın olursa kendi kümesiyle iyi eşleştiğini ve komşu kümelerle eşleşmediğini gösterir. Bu anlamda komşu kümelerle eşleşmemesi, kümenin iyi bir şekilde ayrıştığını gösterir. Değer düşük veya negatifse, kümeleme yapılandırmasında çok fazla veya çok az küme kullanıldığı düşünülebilir. Calinski Harabasz ve Davies Bouldin indeksleri de aynı amaca hizmet etmektedir. Calinski yönteminde sonuç ne kadar yüksek çıkarsa, kümelerin o kadar iyi belirlendiği söylenebilir. Davies Bouldin yönteminde ise, düşük değerler, kümelerin daha iyi ayrıldığını göstermektedir. Bu yöntemde alınabilecek en düşük değer sıfırdır.

Bu çerçevede kapsamında büyük veri ile tüketici ve müşterilere hitap etmek için daha rasyonel sonuçlar elde edilmektedir. Elde edilen bu sonuçlar ışığında müşteri bölümlerine yönelik çoklu regresyon analizleri ile ileriye yönelik davranış modelleri tespit edilmeye çalışılmıştır. Söz konusu kümeleme ve çoklu regresyon analizleri Phyton programı ile uygulanmıştır.

## 4. BULGULAR

### 4.1. Verilere Yönelik Bulgular

Her veri tabanı kendine has birtakım veri depolama ve depolama işlemi sırasında da ayrıştırma işlemi eş zamanlı olarak düzenlenmektedir. Bu çalışmaya ait veri tabanı dört bölümden oluşmaktadır. Bu bölümlere şu şekildedir; siparişe ilişkin veriler (order base data), siparişe ilişkin detaylı veriler (order detail data), ürüne ilişkin veriler (product base data), iletişim ile ilgili veriler (contact base data).

Bu çalışmada bahsedilen veriler "Text" formatında Phyton programına analiz edilmek üzere aktarılmıştır. Bu sayede veri boyutu yaklaşık 300 GB'a düşürülmüştür. Bununla birlikte verilerin ilgili bölümlerde birbirlerini bağlayan otomatik atanmış numaralar bulunmaktadır. Bu numaralar sayesinde ürün ile ilgili bölümde siparişe ait ürün bilgilerini görebilirken, siparişe ait diğer bilgiler (sipariş zamanı, sipariş tutarı, sipariş kanalı vb.) sipariş detay bölümünde bulunabilmektedir. Bu farklı bölümler arasında geçişler de bu siparişleri birbirine bağlayan numaralar sayesinde yapılabilmektedir. Veri tabanında örnek sipariş verilerinin nasıl gözüktüğü ile ilgili ekran görüntüsü aşağıda paylaşılmıştır.

Şekil 1: Veri Tabanı Üzerindeki Örnek Sipariş Verileri

new_storeno	store_city	store_region	new_productid	drink_type	new_netamt	new_cusamt	new_orddt	new_ordersource
40644	İstanbul	MARMARA	045BAE60- C8F8-4E7B- 99BF- 6E8A1A3AB228	"10"" Orta"	49,89	53,89	2018-10-07 00:00:00.000	Web
42800	İstanbul	MARMARA	045BAE60- C8F8-4E7B- 99BF- 6E8A1A3AB228	"10"" Orta"	44,44	47,99	2018-10-02 00:00:00.000	Web
40636	Adana	AKDENİZ	191371AD- 5D42-4D29- A444- 2D29EF0243DA	Adet	70,45	76,09	2018-10-15 00:00:00.000	Phone
42873	İstanbul	MARMARA	191371AD- 5D42-4D29- A444- 2D29EF0243DA	Adet	53,61	57,89	2018-10-21 00:00:00.000	Web
40549	Antalya	AKDENİZ	03589A77- 7394-4B44- ACF1- 4B7058E8CD4A	"13"" Buyuk"	80,90	87,40	2018-10-09 00:00:00.000	Web

Bu tarz büyük veri çalışmalarında tekrar eden numaralar, bu numaraların uzunluğu, kullanılmayacak birtakım veriler çalışmanın boyutunu arttırmaktadır. Bu nedenle öncelikle otomatik olarak atanmış hücre uzunlukları ve kod numaraları yeniden yapılandırılmıştır. Şekil 1’de, ürün numarası ve sipariş tarihi gibi hücrelerin otomatik olarak uzunlukları ve içeriklerini doldurmak için atanmış harf ve rakamlar yeniden düzenlenmiştir. Bazı veriler aynı tablo içerisinde birden fazla kez tekrar etmiştir. Bu tekrar eden hücreler, ilgili veri setinden silinmiştir.

Esas olarak bu bölümde hedef, veriyi kullanılacak programın anlayabileceği hale getirmektir. Makine öğrenme algoritmaları doğrudan kategorik veriler üzerinde çalışmamaktadır. Bu yüzden verilerin sayısal verilere dönüştürülmesi gerekmektedir. Bununla ilgili örnek olarak siparişlerin geldiği gün incelenebilir. Matematiksel bir fonksiyon için “Üç Ocak Çarşamba” gününün bir anlamı yoktur. Bununla birlikte o günü, hafta sonu mu değil mi ya da haftanın üçüncü günü mü şeklinde ifade edebiliriz. Şekil 2’de gösterilen örnekte “one hot encoding” metodu kullanılarak haftanın günleri birer kolona çevrilmiştir. Bu metod, kategorik değişkenlerin ikili (binary) olarak temsil edilmesi anlamına gelmektedir. Örneğimizde ise haftanın ilgili gününde sipariş verilip verilmediğini belirtmek için de 1 ve 0 kullanılmıştır.

Şekil 2: Günlere Göre Sipariş Durum Örneği

Contactid	Newcusamt	cus_amt_mean	cus_amt_count	weekday_0	weekday_1	weekday_2	weekday_3	weekday_4	weekday_5	weekday_6
1	119.97	39.99	3	0	1	0	0	0	0	0
2	55.98	27.99	2	0	0	0	0	0	1	0
3	47.98	23.99	2	0	1	0	0	0	0	0
4	30.00	30.00	1	0	0	0	0	0	0	1
5	144.96	36.24	4	0	0	0	0	0	1	0

Veri işlenmesi aynı zamanda verinin dönüştürülmesi işlemidir. Şekil 2’de olduğu gibi tarih özelliği ile yapılmak istenen analizler belli olduktan sonra mevcut özelliğin anlamlandırılması için veri setine ilave hafta içi ve hafta sonu eklemeleri yapılmıştır. Buna benzer diğer bir çalışma sipariş zamanı özelliğinde kullanılmıştır. Bu noktada sipariş saatinin öğle-akşam-gece olarak anlamlandırılması için veri setine ilave sütunlar eklenmiştir.

Veri tabanında özelliklere ait alt kategorilerin doğru yapılmadığı fark edilmiştir. Örnek olarak ürün kategorileri bölümünde, yan ürün ve tavuk ürünü bir arada kullanılmıştır. Oysa tavuk ürünü yan ürüne ait bir alt sınıftır. Bunun dışında hamur tipleri ve boyları ile yan ürünlere ait parça sayıları aynı özellik altında toplanmıştır. Bu nedenle hamur tipleri ve hamur boyları ayrı bir özellik olarak ayrıştırılmıştır.

Bir diğer önemli temizlik ise veri seti içerisinde özellik olarak bulunan, ancak ilgili veri kaynağında bu veri tutulmadığı için boş olarak gelen hücrelerin temizlenmesidir. Tablo 1’de gösterilen örnekte olduğu gibi, verinin boş olduğunu dört harften oluşan NULL kelimesiyle belirtmesi bile verinin boyutunu ciddi oranda arttırdığı için bu tarz kullanılmayan veriler, veri setinden çıkartılmıştır.

**Tablo 1: Boş Hücrelerin Örnek Görüntüsü**

new_orderdateandnumber	new_campaigncategory	new_campaignname	new_daytimes	new_flavorscore
2018-05-22#59	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL

Veriyi incelemek, verinin doğruluğundan emin olmak ve kullanılacak algoritmaları seçmeden önce fikir sahip olmak için önemlidir. Ayrıca veri içerisindeki “outlier” başka bir ifadeyle aykırı alanları bulmak için de önemli bir yöntemdir. Şekil 3’de gösterilen örnek veri setinde “Elliptic Envelope” yönteminin “outlier” aykırı olarak işaretlediği örnekler bulunmaktadır. Buradaki müşterilerin veri setinde diğer müşterilere göre azınlık durumunda olduğu gösterilmektedir.

**Şekil 3: Elliptic Envelope Yöntemi**

	TotalOrderTyp	TotalWeb	totalNetAmt	AvgNetAmt	CountOrder	mentMethc	TotalPizza	Totalcecek	TotalSides	totalVegiPiz	Recency	isOutlier
count	94895	94895	94895	94895	94895	94895	94895	94895	94895	94895	94895	94895
mean	0,382686	1,723421	103,2517	35,46422	3,022456	1,391823	4,138911	1,213425	1,423289	0,234659	148,7277	0,799989
std	2,286081	4,081807	182,1432	17,67702	4,826751	3,033818	6,622295	10,4978	4,05024	1,229989	107,9499	0,600017
min	0	0	0,31	0,062	1	0	0	0	0	0	1	-1
25%	0	0	31,3	25,18146	1	0	1	0	0	0	51	1
50%	0	1	50,47	32,36857	1	1	2	1	0	0	133	1
75%	0	2	107,685	41,56513	3	1	4	1	1	0	238	1
max	270	271	24913,32	1111,11	293	157	364	3126	369	67	365	1

Örneğin, “OrderTypeC” gel al ve masaya servis kolonunda %25 - %50 - %75 değerleri için sıfır yazmaktadır. Bir başka ifade ile, ilgili kolon içerisindeki değerlerin kalan %25’lik kısımda arttığını görmekteyiz. “TotalVegiPizza” için aynı kontrol yapıldığında, ürünün ortalama 0,2 adet satın alındığı ve standart sapma değerinin 1,2 olduğu görülmektedir. Müşteriler tarafından çok fazla tercih edilmeyen bu ürünün %25-%50-%75 değerleri sıfırdır. Sadece kalan %25’lik dilimde tercih edildiğini göstermektedir. Müşterilere ait sipariş adetleri incelendiğinde, ortalama 1,3 adet sipariş ve %75’lik dilime kadar 1 sipariş verildiği görülmektedir. Bir başka ifade ile, müşterilerin en az yarısının sadece bir kez sipariş vermiş oldukları söylenebilir.

Bu yöntem ayrıca “outlier” aykırı olarak aynı işlemi tekrar etmiştir. Bunun neticesinde ortalama %8 civarında aykırılıkların olduğunu tespit etmiştir. Bunu şu şekilde inceleyebiliriz: Örneğin; ortalama sepet tutarının 35,46 TL olduğu bir ürüne en fazla 1.111 TL ödemenin yapılması algoritma tarafından aykırı olarak belirlenmiştir. Benzer şekilde yapılan alışverişlerin toplam tutarı ortalama 103 TL olan bir örnek grubunda, Tablo 2’de gösterildiği gibi en fazla toplam 24.913 TL ödeme yapıldığı tespit edilmiştir. Bu ve benzer sonuçlar, azınlık olarak belirlendikten sonra aykırı siparişler daha detaylı incelenmiştir.

**Tablo 2: Aykırılıklara Ait Örnekler**

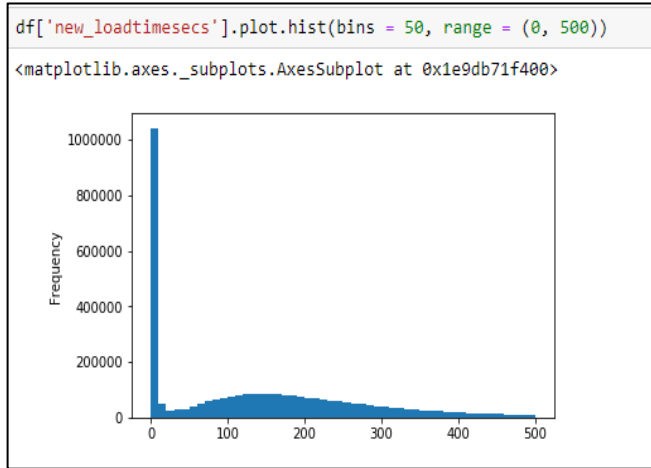
CustomerID	TotalOrderTypeC	TotalWeb	TotalNetAmt	AvgNetAmt	CountOrder	Recency	isOutlier
37050	144	51	24913,32	154,7411	161	27	-1
3770300	6	17	9426,75	428,4886	22	48	-1
3817900	195	226	6656,18	29,06629	229	6	-1
1717200	1	0	6434,49	40,98401	157	1	-1
2045250	270	271	5565,98	18,99652	293	38	-1
1886900	150	71	5256,65	31,85848	165	1	-1
14750	197	198	4502,4	22,07059	204	4	-1

Bu siparişlerin ortalamasının çok üzerinde olması, bunun gerçek bir müşteri davranışı olmayacağı anlamına gelmemektedir. Bu nedenle, yanlış karar vermek adına ilgili siparişlerin verildiği şubeler rastgele aranmış ve kontrol edilmiştir. Alınan geri

bildirimler sonucunda bu ve benzeri siparişlerin genelde anlaşmalı okul, iş yeri veya markanın başış şeklinde yaptığı toplu gönderimler olduğu tespit edilmiştir.

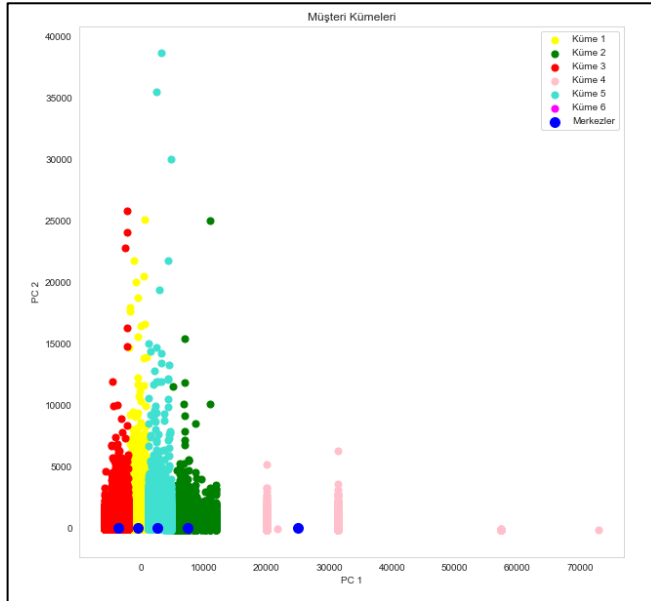
Şekil 4’de gösterilen diğer bir örnekte ise “Loadtimesecs” ürün yükleme süreleri kolonunun veri içerisinde ne kadar olduğuna bakarak, bu özelliği algoritma içerisinde kullanılmaması gerektiği tespit edilmiştir. Çünkü içerisindeki verilerin büyük bir çoğunluğu sıfır ya da sıfıra yakın bir değer girildiği ya da bu şekilde işlendiği için bu durum gerçek değerlerin azınlıkta kalmasına sebep olmuştur. Sonuç olarak, bu verinin doğru bir şekilde tutulmadığını ya da hesaplanmadığını göstermektedir.

**Şekil 4: Loadtimesecs Özelliğinin Veri Tabanındaki Dağılımı**



Verileri incelerken bir diğer yöntem de görselleştirmedir. Genel olarak PCA metodu kullanılır. Örneğin, veri setindeki özellikleri PCA ile iki boyuta indirgedikten sonra algoritma aracılığı ile ayrılan kümeleri Şekil 5’de olduğu gibi görselleştirilebilir. PCA metodu verideki özelliklerin iki boyuta indirgenmesine yardımcı olur. İki boyuta indirgenen verileri incelemek ve görselleştirmek, çok daha kolaydır.

**Şekil 5: PCA Analiz Örneği**



#### 4.2. Kümeleme Analizine Yönelik Bulgular

Model oluşturulurken dikkat edilmesi gereken birçok faktör vardır. Bunlar çoğu zaman makinanın kendi kendine belirleyemediği parametrelerdir. Gözlemlere ve sonuçlara dayanarak bunlar üzerinde oynamalar yapılabilir. İlk yapılması gereken, hangi algoritmanın kullanılacağına karar verilmesidir. Bu amaçla en yaygın olarak kullanılan yöntem, Sillhoute (Rousseeuw, 1987)

katsayıdır. Bunun dışında Calinski Harabasz (Calinski, Harabasz, 1974) indeksi ve Davies Bouldin (Davies, Bouldin, 1979) indeksi de kullanılmaktadır.

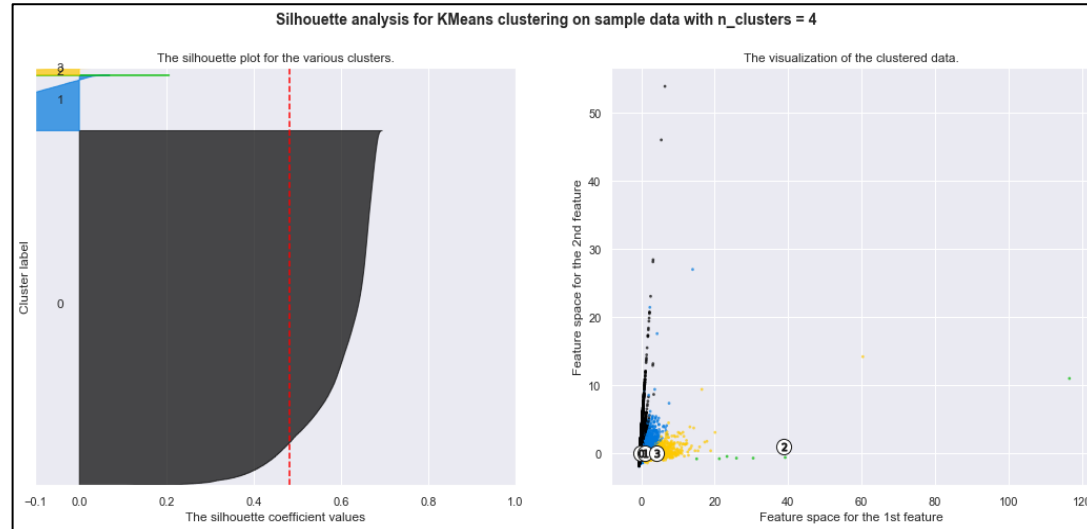
Şekil 6'da iki farklı kümeleme algoritması ile üç farklı küme sayısına bölünmüş kümeler karşılaştırılmıştır.

#### Şekil 6: Karşılaştırmalı Kümeleme Algoritmaları

4 Küme	6 Küme	8 Küme
Counter({0: 68452, 1: 10666, 3: 1405, 2: 7})	Counter({2: 62829, 0: 13640, 5: 3508, 1: 546, 3: 5, 4: 2})	Counter({1: 62156, 4: 13635, 2: 3012, 7: 722, 0: 537, 5: 462, 3: 5, 6: 1})
<b>Silhouette K Means:</b> 0.4829907771272668	<b>Silhouette K Means:</b> 0.36055848530667245	<b>Silhouette K Means:</b> 0.3392441652448518
<b>Calinski</b> 9.204.889.633.914.360	<b>Calinski</b> 7.572.199.399.825.330	<b>Calinski</b> 6.180.182.980.254.140
<b>Davies_Bouldin</b> 19.768.011.575.979.600	<b>Davies_Bouldin</b> 20.917.696.527.922.100	<b>Davies_Bouldin</b> 19.572.664.328.115.700
<b>Silhouette Gaussian:</b> 0.37198886912729395	<b>Silhouette Gaussian:</b> -0.02650809286035061	<b>Silhouette Gaussian:</b> 0.07850418466811937
<b>Calinski Gaussian</b> 7.148.471.931.445.330	<b>Calinski Gaussian</b> 4.868.359.182.836.330	<b>Calinski Gaussian</b> 43.786.746.745.761.300
<b>Davies_Bouldin Gaussian</b> 24.475.118.148.340.400	<b>Davies_Bouldin Gaussian</b> 35.534.791.471.526.600	<b>Davies_Bouldin Gaussian</b> 27.859.234.809.572.800

Silhouette, Calinsky ve Davies Bouldin değerleri şekil üzerinde gösterilmiştir. Öncelikle K Means algoritması, sekizli kümede Calinski değeri hariç tüm incelemelerde daha iyi sonuç vermiştir. Her üç değer de en optimum seviyeye ulaştığı küme sayısı ise dörtlü kümedir. Bu kümede, Silhouette değeri diğer altılı ve sekizli kümelere kıyasla en yüksek değere ulaşmıştır. Aynı durum Calinski ve Davies değerleri için de geçerlidir. Bu doğrultuda çalışmada seçilen K Means dörtlü kümesinin Silhouette görseli Resim 1'de paylaşılmıştır.

#### Resim 1: K Means 4'lü Kümeleme Silhouette Analizi

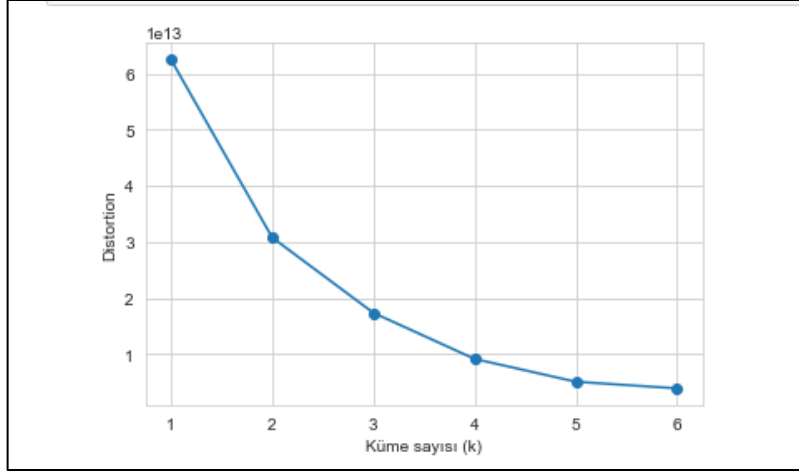


Resim 1'de soldaki görselde kırmızı kesik çizgi Silhouette katsayısını göstermektedir. Bu değer 0,48 olarak çıkmıştır. Sağ taraftaki görselde ise kümelerin nasıl dağıldığı görselleştirilmiştir. Yeşil renkli kümenin aykırı olduğu, tek ekseninde ince bir çizgi halinde seyir etmesiyle de belli olmaktadır.

Yapılan analizler sonucunda K Means 4'lü küme sonucu en optimum sonucu verse de küme sayısının belirlenmesinde bir diğer önemli metod ise Elbow'dur (Kordinariya ve Makwana, 2013). Elbow metodunda grafiğin eğiminin büyük ölçüde değiştiği nokta küme sayısı için baz alınır. Şekil 7'de Elbow metodu ile veri setimizde oluşan grafik çıktısı gösterilmiştir.



Şekil 7: Elbow Yöntemine Ait Grafik



Bu metoda göre küme sayısı dört olarak seçilebilir. Bu yöntemde bir kesinlik yoktur ve yoruma açıktır. Modelin kırılma noktası, kesin olarak dört olarak belirtilemeye de sonuçları gözlemleyerek ve yorumlayarak ilerlemek gerektiği söylenebilir. Neden böyle olduğunu örnekle açıklamak gerekirse; bir veri setinde müşteriler ve bu müşterilerin harcamaları olduğu varsayıldığında, müşterileri sadece harcamalarına göre kümelemek için ideal küme sayısı diye bir sayı yoktur. Burada müşteriler çok harcama yapanlar ve az harcama yapanlar olarak iki kümeye de ayrılabilir ya da çok harcama yapanlar, orta harcama yapanlar ve az harcama yapanlar olarak üç kümeye de ayrılabilir. Bu çalışmada kümeleme sonuçlarının değerlendirilmesi açısından birbirinden farklı veri setleri kullanılmış ve kümeler arasında ayırt edici özelliği bulunmayan özelliklerinin elenmesi yöntemiyle farklı kümeleme çalışmaları test edilmiştir. Aynı zamanda tüm çalışmalarda dördü, altı ve sekizli küme analizleri yapılmıştır.

- Tüm özelliklerin veri setinde bulunduğu, 100.000 adet sipariş verisinin analizi
- Tüm özelliklerin veri setinde bulunduğu, farklı 100.000 adet sipariş verisinin analizi
- İlk veri setindeki farklılaşmayan özelliklerin azaltılarak yeniden analiz edilmesi
- İlk veri setindeki azaltılmış özelliklerin yeniden farklılaşmayan özelliklerin azaltılarak analiz edilmesi
- Büyük kümenin alt kümelere ayrıştırılması

Kümeleme sonuçlarına ait tabloların anlaşılması açısından açıklamalar, Ek 1’de gösterilmiştir. Ayrıca analiz sonuçlarının ve takip eden alt bölümlerdeki tabloların rahat okunabilmesi için aynı özellikteki kümelere aynı harf atanmıştır. Ek 2’de gösterilen tüm özelliklerin bulunduğu bu ilk 100.000 adetlik veri setinin sonuçları incelendiğinde, A, B ve C kümelerinin ayrıştığı ve X kümesinin aykırı olduğu gözlemlenmiştir.

X kümesinin sipariş adetlerinin fazla olması nedeniyle detaylı inceleme yapıldığında, bunun sebebinin şubeler tarafında oluşturulan merkezi içecek gönderimi olduğu anlaşılmıştır. Bu noktada algoritmanın X kümesini diğerlerinden ayırt etmesi ve “outlier” başka bir ifadeyle aykırı olduğunu ortaya çıkarması nedeniyle, anomali ve farklılıkları doğru bir şekilde tespit edebildiğini göstermiştir.

A kümesi incelendiğinde, kümenin, tüm müşterilerin %85’ini oluşturduğu ve bu kümeye ait müşterilerin diğer kümelere kıyasla en az sipariş veren küme olduğu söylenebilir. Buradan yola çıkarak verileri incelenen pizza firmasını, yeni müşteri ağırlıklı ve frekansı düşük müşterilerin tercih ettiği söylenebilir. Bunun sebebinin neler olduğu bu çalışmanın kapsamında değildir. Bununla birlikte toplam cironun %50’si bu küme tarafından gerçekleştirilmektedir. İnternette sipariş verme oranı en düşük olan bu küme, aynı zamanda yan ürün tercihi olarak da en son sırada yer almaktadır. Çift katlı taban gibi özellikli ürünleri tercih etmeyen bu kümenin pizza ağırlıklı alışveriş yaptığı söylenebilir. Bu noktada hem frekansının düşük olması hem de pizza oranının yüksek olması, bu kümenin promosyon veya kampanya zamanlarında alışveriş yapan ve fiyata duyarlı olan bir müşteri grubu olma ihtimalini arttırmaktadır. Karışık pizza oranının diğer kümelerin üzerinde olması, pizza çeşitlerine hakim olmadıklarını gösteriyor olabilir. Nakit ödeme oranının yüksek olması, bu kümenin daha cazip kampanyaları tercih ettiklerini gösteriyor olabilir.

A kümesi, tüm müşterilerin %85’ini oluşturduğu için K-Means algoritmasında yeniden alt kümelere ayrılmak üzere analiz edilmiştir. Bu doğrultuda Ek 3’de gösterildiği üzere A kümesinin A1 ve A2 olarak iki alt kümeye ayrıldığı gözlenmiştir.

A kümesinin yeniden analiz edilmesi sonrasında A1 kümesinin müşterilerin %64'ünü ve cironun %75'ini oluşturduğu gözlenmiştir. A2 kümesi ise müşterilerin %36'sını ve cironun %25'ini oluşturmaktadır. Bu iki kümenin temel ayrıştığı özellikler, Frekans – Son Sipariş Zamanı ve Toplam Pizza İçindeki Oranı olmuştur. Bu noktada A1 kümesinin daha fazla sipariş veren ve pizza miktarı daha fazla olan küme olduğu söylenebilir. A2 kümesinin sene başında sipariş vermiş ve tekrar alışveriş yapmamış olduğu gözlenmektedir.

B kümesi, müşterilerin %13'ünü ve toplam cironun %35'ini oluşturmaktadır. A kümesinden sonra hem müşteri sayısı hem de ciro payı olarak en büyük ikinci kümedir. Yıl içerisinde ortalama yedi sipariş veren bu küme, internet siparişlerini en fazla kullanan ikinci kümedir. Yan ürün tercihleri A numaralı gruba göre daha yüksek olmakla birlikte yaklaşık olarak her iki siparişin birinde yan ürün sipariş verdikleri söylenebilir. Akşam siparişleri diğer kümelere göre en yüksek olan kümedir.

C kümesi, müşterilerin sadece %1'ini oluşturmasına rağmen toplam cironun %13'ünü oluşturmaktadır. Yıl içerisinde 20 adet sipariş ile en fazla sipariş veren kümeyi oluşturmaktadır. Yan ürün satın alma oranı en yüksek olan bu kümenin çift katlı taban gibi özellikli ürünleri en çok tüketen küme olduğunu söyleyebiliriz. Bununla birlikte etsiz pizzaların da %7 gibi bir oranla tüm kümeler arasında en yüksek değere sahip olduğu dikkat çekmektedir. Sipariş saatine göre incelendiğinde ise, gece siparişinin en yüksek olduğu söylenebilir. Sipariş sıklığı ve sadece pizza tercih etmeyip diğer özellikli ve yan ürünleri de kullanıyor olmaları nedeniyle bu kümenin pizzayı seven, ürün algısı ve dışardan yeme alışkanlığı yüksek, aynı zamanda yeniliklere açık bir küme olduğu söylenebilir. İnternet kullanımının ve aynı zamanda gece siparişlerinin de en yüksek oranda olması, bu kümeyle ilgili pazarlama stratejilerinde üzerinde durulması gereken bir nokta olarak dikkat çekmektedir.

Bununla birlikte çalışmada kullanılan toplam 52 özellikten bazılarının diğer kümeler arasında anlamlı bir fark yaratmadığı tespit edilmiştir. Özellikle coğrafik analizlerde kullanılması için eklenen şehir ve bölge kırılımı, diğer kümelerle benzer dağılım göstermektedir. Ayrıca evlere servis, gel al ve masaya servisi ortalama sepet tutarlarının 15'er TL'lik dilimlere ayrılarak bu fiyat aralıklarındaki dağılımlar analiz edilmek istenmiş olup analiz sonuçlarında anlamlı bir fark bulunamamıştır. Bunun gibi farklılık bulunmayan diğer özellikler de bir sonraki çalışmadan çıkartılmıştır.

Farklı algoritma ve küme sayıları ile yapılan çalışmalarda ve azaltılmış özelliklerin test edildiği analizler doğrultusunda müşterilerin üç temel küme etrafında ayrıştığı tespit edilmiştir. Bunlar A, B ve C kümeleridir. Küme sayısı arttıkça bu kümelere ait analizlerde yeni kümelerin değil, mevcut kümelere ait alt kümelerin olduğu gözlenmiştir. Bununla birlikte küme sayısı arttıkça alt kümeler bir veya iki özellik tarafından ayrıldığı için genel olarak benzerliklerini korumakla birlikte bu kümelere ait ayrıştırıcı özellikler yetersiz kalmaktadır. Bu değerlendirme çalışmanın başında yapılan Sillhoutte, Calinski Harabasz ve Davies Bouldin indeksleri ile de paralellik göstermiştir. Bu nedenle pazarlama stratejileri oluşturma konusunda farklılıklarını koruyan temel A, B ve C kümeleri, çoklu regresyon analizlerinde kullanılmak üzere seçilmiştir. Ayrıca A kümesinin büyüklüğü nedeniyle ve kümeyi daha iyi tanımlayabilmek adına A1 ve A2 alt kümeleri kullanılması kararlaştırılmıştır. Seçilen küme ve özellikler Tablo 3'de gösterilmiştir.

**Tablo 3: Seçilen Müşteri Kümeleri**

	MÜŞTERİ KÜMELERİ			
	A1	A2	B	C
Toplam Müşteri İçindeki Oranı	64%	36%	13%	1,0%
Ciro İçindeki Oranı	75%	25%	35%	13%
Frekans	2,2	1,4	7	20
Recency	74	285	59	27
ATP / Ort Sepet Tutarı	36,87	33,30	35,27	34,35
Web Sipariş Oranı	54,8%	47,3%	65,3%	67,0%
Toplam Pizza İçindeki Oranı	74,2%	25,8%	35,8%	12,3%
Yan Ürün Oranı	43,1%	43,6%	49,0%	62,0%
Çift Katlı Taban Oranı	1,6%	1,5%	2,7%	5,1%
Etsiz Pizza Oranı	4,8%	4,5%	5,5%	7,6%
Küçük Pizza Oranı	16,0%	13,0%	16,1%	18,9%
Karışık Pizza Oranı	15,1%	17,7%	11,4%	10,9%
Çok Malzemeli Pizza Oranı	20,0%	23,4%	19,8%	17,6%
Akşam Sipariş Oranı	59,1%	59,0%	59,2%	55,3%
Nakit Ödeme Oranı	55,4%	58,3%	49,8%	50,6%
Gel Al ve Masaya Servis Oranı	13,4%	11,3%	13,0%	10,9%

Evlere Servis Oranı	86,6%	88,7%	87,0%	89,1%
---------------------	-------	-------	-------	-------

### 4.3. Çoklu Regresyon Analizine Yönelik Bulgular

Her bir küme için şu bağımlı değişkenler test edilmiştir: Frekans, ortalama sepet tutarı, evlere servis oranı, gel al ve masaya servis oranı ve pizza oranı.

#### 4.3.1. Bağımlı Değişken: Frekans / Sipariş Sıklığı Olmak Üzere Çoklu Regresyon Analizi

Müşterilerin çoklu ve tekil olmak üzere tüm sipariş adetlerinin analizi ve bu doğrultuda frekans diğer bir ifade ile sipariş sıklığı ile ilgili çoklu regresyon sonuçları incelenmiştir.

#### A1 Kümesi için Çoklu Regresyon Analizi

Değişkenlere ait VIF değerleri Şekil 8'de gösterilmiştir.

Şekil 8: A1 kümesi VIF Değerleri

VIF Factor	features
0	2.71 Recency
1	2.19 TotalWeb
2	1.37 TotalSides
3	1.03 TotalDublex
4	1.14 TotalVegiPizza
5	1.30 TotalKucukPizza
6	1.60 TotalKarisikPizza
7	1.54 TotalBolMalzemosPizza
8	2.55 TotalAksam
9	3.00 TotalPaymentMethodType26
10	7.97 TotalPizza
11	6.08 AvgNetAmt
12	9.41 TotalOrderTypeD
13	2.57 TotalOrderTypeC

CountOrder (frekans) başka bir ifade ile sipariş sıklığı; TotalOrderTypeD (evlere servis) ve TotalOrderTypeC (gel al ve masaya servis) siparişlerinin toplamı olduğu için, bu iki bağımsız değişkene ait VIF değeri 10'un altında bir değer olsa da Şekil 9'da paylaşılan regresyon analizinde 1'e eşit çıktığı görülmektedir. Ayrıca R<sup>2</sup> değeri de 1 çıkmıştır.

Şekil 9: A1 Kümesi İçin Bağımlı Değişkenin Frekans Olduğu Çoklu Regresyon Analizi

OLS Regression Results						
Dep. Variable:	CountOrder	R-squared (uncentered):	1.000			
Model:	OLS	Adj. R-squared (uncentered):	1.000			
Method:	Least Squares	F-statistic:	6.991e+32			
Date:	Mon, 25 May 2020	Prob (F-statistic):	0.00			
Time:	08:49:53	Log-Likelihood:	1.1589e+06			
No. Observations:	35883	AIC:	-2.318e+06			
Df Residuals:	35869	BIC:	-2.318e+06			
Df Model:	14					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Recency	1.392e-17	9.5e-20	146.588	0.000	1.37e-17	1.41e-17
TotalWeb	7.654e-16	2.34e-17	32.723	0.000	7.2e-16	8.11e-16
TotalSides	1.214e-16	1.47e-17	8.256	0.000	9.26e-17	1.5e-16
TotalDublex	-7.841e-16	8.12e-17	-9.660	0.000	-9.43e-16	-6.25e-16
TotalVegiPizza	3.521e-16	4.4e-17	8.006	0.000	2.66e-16	4.38e-16
TotalKucukPizza	4.361e-16	2.16e-17	20.229	0.000	3.94e-16	4.78e-16
TotalKaristikPizza	3.023e-16	2.28e-17	13.241	0.000	2.58e-16	3.47e-16
TotalBolMalzemosPizza	-8.5e-17	2.12e-17	-4.010	0.000	-1.27e-16	-4.34e-17
TotalAksam	3.66e-16	2.22e-17	16.496	0.000	3.23e-16	4.1e-16
TotalPaymentMethodType26	1.687e-16	2.29e-17	7.358	0.000	1.24e-16	2.14e-16
TotalPizza	-2.689e-17	1.92e-17	-1.401	0.161	-6.45e-17	1.07e-17
AvgNetAmt	-7.36e-17	7.62e-19	-96.543	0.000	-7.51e-17	-7.21e-17
TotalOrderTypeD	1.0000	3.36e-17	2.97e+16	0.000	1.000	1.000
TotalOrderTypeC	1.0000	4.51e-17	2.22e+16	0.000	1.000	1.000
Omnibus:	1797.847	Durbin-Watson:	1.991			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3131.666			
Skew:	0.405	Prob(JB):	0.00			
Kurtosis:	4.200	Cond. No.	1.42e+03			

Diğer kümelere ait regresyon sonuçlarında TotalOrderTypeD başka bir anlatımla, evlere servis sipariş değeri, çoğunlukla 10'un üzerinde çıktığı için bu değişken, bağımsız değişkenlerden çıkartılmıştır ve bağımlı değişken olarak kullanılmıştır. Bu değişkenin yerine TotalOrderTypeC diğer bir deyişle, gel al ve masaya servis siparişleri bağımsız değişken olarak kullanılmıştır. Bu doğrultuda yeni VIF sonuçları Şekil 10'da gösterilmiştir.

Şekil 10: A1 kümesi VIF Değerleri

VIF Factor	features
0	2.51 Recency
1	1.91 TotalWeb
2	1.35 TotalSides
3	1.03 TotalDublex
4	1.14 TotalVegiPizza
5	1.29 TotalKucukPizza
6	1.56 TotalKaristikPizza
7	1.53 TotalBolMalzemosPizza
8	2.25 TotalAksam
9	2.29 TotalPaymentMethodType26
10	7.55 TotalPizza
11	5.88 AvgNetAmt
12	1.32 TotalOrderTypeC

Her deęişkenin faktör deęeri istenilen sınırlar içinde kaldığı için bir sonraki aşama olan çoklu regresyon sonuçları Şekil 11’de incelenmiştir.

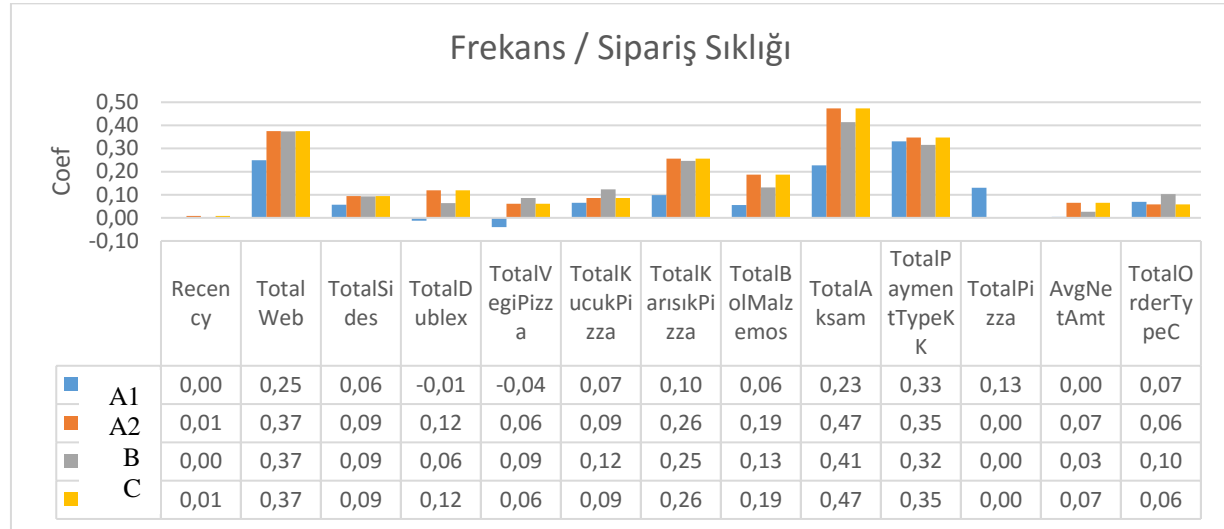
**Şekil 11: A1 Kümesi İçin Yeniden Oluşturulan Çoklu Regresyon Analizi**

OLS Regression Results						
Dep. Variable:	CountOrder	R-squared (uncentered):	0.910			
Model:	OLS	Adj. R-squared (uncentered):	0.910			
Method:	Least Squares	F-statistic:	2.779e+04			
Date:	Wed, 27 May 2020	Prob (F-statistic):	0.00			
Time:	11:28:46	Log-Likelihood:	-14019.			
No. Observations:	35883	AIC:	2.806e+04			
Df Residuals:	35870	BIC:	2.817e+04			
Df Model:	13					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Recency	0.0008	1.43e-05	55.392	0.000	0.001	0.001
TotalWeb	0.2489	0.003	72.567	0.000	0.242	0.256
TotalSides	0.0572	0.002	24.973	0.000	0.053	0.062
TotalDublex	-0.0121	0.013	-0.952	0.341	-0.037	0.013
TotalVegiPizza	-0.0397	0.007	-5.748	0.000	-0.053	-0.026
TotalKucukPizza	0.0658	0.003	19.556	0.000	0.059	0.072
TotalKarisikPizza	0.0980	0.004	27.640	0.000	0.091	0.105
TotalBolMalzemosPizza	0.0561	0.003	16.908	0.000	0.050	0.063
TotalAksam	0.2273	0.003	69.520	0.000	0.221	0.234
TotalPaymentMethodType26	0.3309	0.003	105.127	0.000	0.325	0.337
TotalPizza	0.1300	0.003	44.328	0.000	0.124	0.136
AvgNetAmt	0.0041	0.000	34.772	0.000	0.004	0.004
TotalOrderTypeC	0.0692	0.005	13.600	0.000	0.059	0.079
Omnibus:	1483.128	Durbin-Watson:	1.947			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2892.746			
Skew:	0.310	Prob(JB):	0.00			
Kurtosis:	4.245	Cond. No.	1.42e+03			

A1 kümesi için yapılan regresyon çalışmasında  $R^2$  deęeri 0,91 olarak çıkmıştır. Deęişkenler içerisinde gözlenen en yüksek deęerler, ödeme yöntemi, web siparişi ve akşam siparişi olmuştur. Frekans ile ters yönlü ilişkide olan deęişkenler ise çift katlı taban ve etsiz pizza siparişleri olmuştur.

A1 kümesi özelinde paylaşılan çoklu regresyon analiz süreci dięer kümelerde de aynı şekilde yapılmış olup bundan sonraki bölümlerde her küme için aynı analizleri paylaşmak yerine tüm kümeler üzerindeki sonuçları toplu şekilde gösterilecektir.

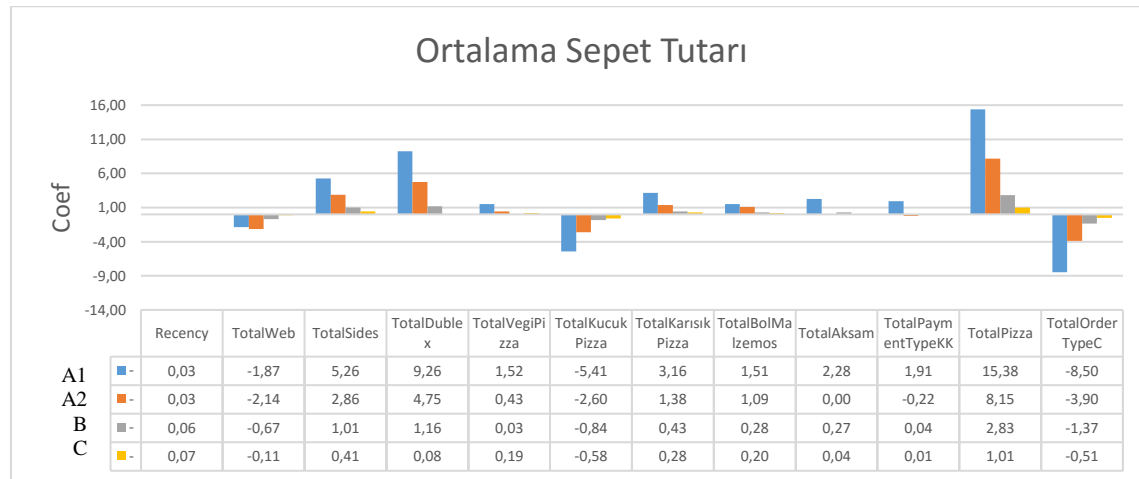
Bağımsız deęişkenlere ait deęerlerin tüm kümeler üzerindeki dağılımı Şekil 12’de gösterilmiştir. En yüksek deęerlerin akşam ve internet siparişleri ile kredi kartıyla ödeme yönteminde olduğu gözlenmiştir.

**Şekil 12: Bağımlı Değişken: Frekans / Sipariş Sıklığı Üzere Çoklu Regresyon Analizi**

Kümeler üzerindeki dağılımına baktığımızda A1 kümesinin etsiz pizza seçimlerinde ters korelasyon ile hareket ettiği gözlenmektedir. Bununla beraber çift katlı taban gibi özellikli ürünlerde de aynı eğilim bulunmaktadır. A2 ve C kümelerinin ise akşam siparişlerinde en yüksek artış yapma potansiyelinin olduğu gözlenmektedir. B kümesinin ayırt edici bağımsız değişken sonucu ise küçük boy pizza ve etsiz pizza artışlarında gözlenmiştir. Ayrıca gel al ve masaya servis siparişlerinde en fazla artış gösteren kümedir.

#### 4.3.2. Bağımlı Değişken: Ortalama Sepet Tutarı Üzere Çoklu Regresyon Analizi

Bağımsız değişkenlere ait değerlerin kümeler üzerindeki dağılımı Şekil 13'de gösterilmiştir. En yüksek değerlerin pizza adedi ve çift katlı taban seçeneğinde olduğu gözlenmiştir. Bununla birlikte üç değişkende negatif yönlü hareket olmuştur. Bu üç değişken, internet siparişleri, küçük pizza ve gel al ve masaya servis siparişleridir.

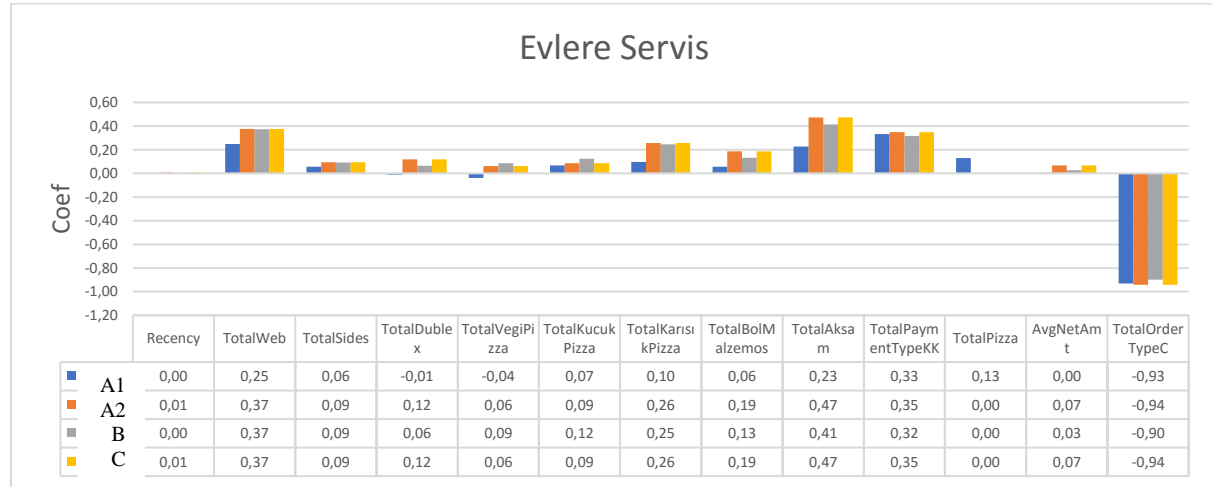
**Şekil 13: Bağımlı Değişken: Ortalama Sepet Tutarı Üzere Çoklu Regresyon Analizi**

Kümeler üzerindeki dağılımına baktığımızda A1 kümesinin internet siparişleri hariç tüm bağımsız değişkenler üzerinde en yüksek değere ulaştığı söylenebilir. En az sipariş veren küme olmasına rağmen fiyat ile ilgili bariyerlerin kaldırılması durumunda hem ürün hem de yan ürün tüketiminin en yüksek seviyede artacağı gözlenmiştir. Bu doğrultuda sadakat programları veya tekrar siparişlere ekstra indirim sağlanması gibi uygulamaların A1 kümesini teşvik edeceği düşünülmüştür. Benzer sonuçlar A2 için de söylenebilir.

#### 4.3.3. Bağımlı Değişken: Evlere Servis Siparişleri Olmak Üzere Çoklu Regresyon Analizi

Bağımsız değişkenlere ait değerlerin kümeler üzerindeki dağılımı Şekil 14'da gösterilmiştir. En yüksek değerlerin akşam siparişleri, kredi kartı ile ödeme yöntemi ve internet siparişlerinde olduğu gözlenmiştir. Bununla birlikte gel al ve masaya servis değişkeninde negatif yönlü hareket tespit edilmiştir.

Şekil 14: Bağımlı Değişken: Evlere Servis Siparişleri Olmak Üzere Çoklu Regresyon Analizi

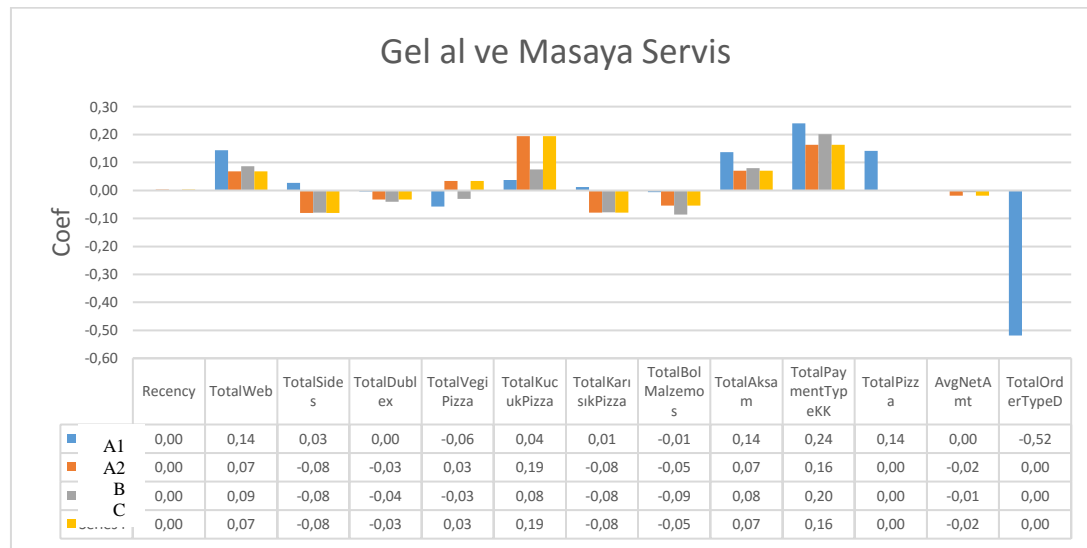


Kümeler üzerindeki dağılımına baktığımızda A1 kümesinin etsiz pizza tercihlerindeki negatif eğilimin bağımlı değişkenin evlere servis olduğu siparişlerde de devam ettiği gözlenmiştir. A2 ve C kümeleri tüm bağımsız değişkenlerde aynı eğilimi göstermektedir. Bu durum diğer çoklu regresyon analizinde de geçerliliğini korumaktadır. Bu nedenle en az sipariş veren A2 kümesinin, en değerli C kümesi ile aynı davranış özellikleri sergiledikleri söylenebilir. B kümesinin küçük pizza ve etsiz pizza özelindeki yüksek değeri bu bağımlı değişken özelinde de devam etmektedir.

#### 4.3.4. Bağımlı Değişken: Gel Al ve Masaya Servis Siparişleri Olmak Üzere Çoklu Regresyon Analizi

Bağımsız değişkenlere ait değerlerin kümeler üzerindeki dağılımı Şekil 15'de gösterilmiştir. En yüksek değerlerin kredi kartı ile ödeme yönteminde gözlenmiştir. Bununla birlikte özellikle etsiz pizzalarda ve küçük boy pizza değişkenlerinde kümeler arasında farklar oluştuğu gözlenmiştir.

Şekil 15: Bağımlı Değişken: Gel Al ve Masaya Servis Siparişleri Olmak Üzere Çoklu Regresyon Analizi

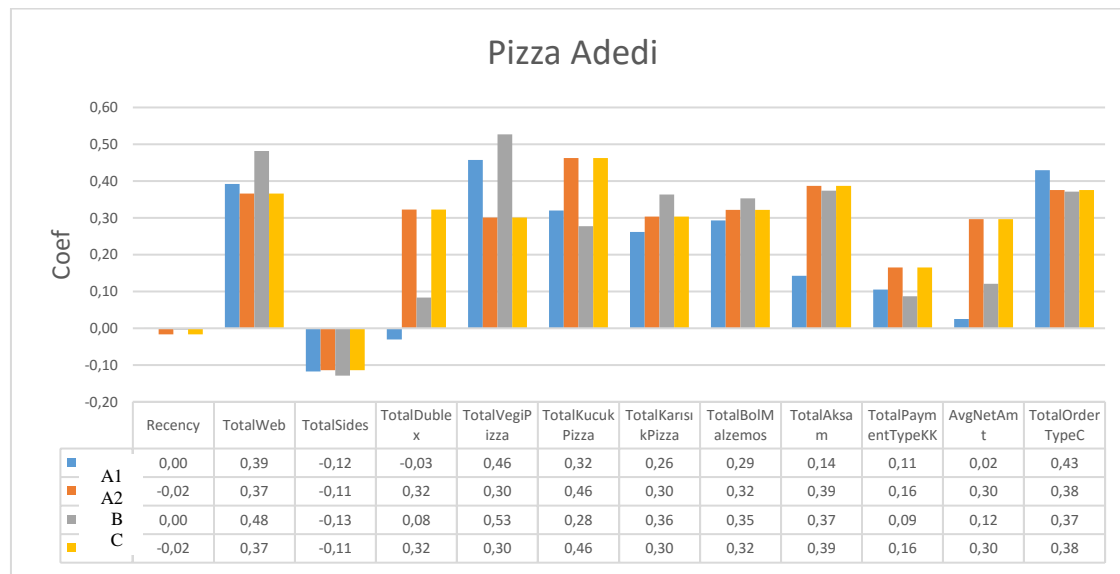


Kümeler üzerindeki dağılımına baktığımızda gel al ve masaya siparişlerde yan ürün kullanımının A1 kümesi hariç diğer tüm kümelerde negatif korelasyon ile hareket ettiği gözlenmiştir. Aynı şekilde özellikli ürün ve çok malzemeli pizzalarda da aynı eğilim devam etmiştir. Bu bağımlı değişken özelinde ortalama sepet tutarının düştüğü ve bu kanalı daha ucuz fiyat nedeniyle tercih edildiği söylenebilir. Bununla birlikte bu satış kanalında şube içerisinde internetten sipariş vererek beklemeden gel al kullanımının da arttığı gözlenmiştir ve akşam saatlerinde bu kanalda sipariş yoğunluğunun yaşanacağı söylenebilir.

#### 4.3.5. Bağımlı Değişken: Pizza Adedi Olmak Üzere Çoklu Regresyon Analizi

Bağımsız değişkenlere ait değerlerin kümeler üzerindeki dağılımı Şekil 16'da gösterilmiştir. Yan ürün hariç hemen hemen tüm değişkenlerde yüksek değerler gözlenmiştir. A1 kümesi çift katlı taban değişkeninde diğer kümelerden farklı yönde ilişki sergilemiştir.

Şekil 16: Bağımlı Değişken: Pizza Adedi Olmak Üzere Çoklu Regresyon Analizi



Kümeler üzerindeki dağılımına baktığımızda ise B kümesinin bağımsız değişkenler üzerindeki dağılımı dikkat çekmektedir. Özellikle internetten sipariş, etsiz pizza siparişi, karışık pizza ve çok malzemeli pizza değişkenlerinde en yüksek değere ulaşmıştır. Bu küme özelinde internete yönelik ve özellikli ürün iletişiminin olduğu tutundurma faaliyetlerinin etkili olacağı düşünülmüştür. A2 ve C kümelerinin paralel hareketleri bu bağımlı değişken özelinde de devam etmekte olup A2 kümesinin potansiyelini bir kez daha ortaya koymaktadır.

## 5. SONUÇ VE ÖNERİLER

Teknolojik gelişmelerin artması ve internetin yaygınlaşmasına paralel olarak, veri madenciliği ve büyük veri yönetimi alanında geliştirilen teknikler, pazarlama faaliyetlerine her türlü veriyi anlık olarak takip etme imkanı sağlamıştır. İşletmeler bilginin ulaşılabilir olması, maliyetinin azalması ve dolayısıyla değerinin artması ve müşteri odaklı pazarlama anlayışının gelişmesiyle birlikte, ilişkide buldukları müşteriler hakkında çeşitli verileri toplamak, depolamak ve bu verileri yararlı bir şekilde bilgiye dönüştürerek stratejiler geliştirmek amacıyla birtakım çalışmalara yönelmişlerdir.

Bu çalışmaları inceleyen veri tabanlı pazarlama sayesinde müşterilere ait verilerden yola çıkılarak etkili stratejiler geliştirilmektedir. İşte bu kapsam dahilinde yapılan uygulama hem büyük veri ile veri tabanlı pazarlama literatüründe farklı bir örnek oluştururken hem de uygulamacılara örnek bir model ve strateji belirleme metodolojisi sunmaktadır.

Pazar bölümlenme ile ilgili literatürde en çok kullanılan ve Kotler'in ortaya atmış olduğu bölümlenme kriterleri kullanılmıştır. Bunlar; demografik, coğrafik, psikografik ve davranışsal kriterlerdir. Veri setinin içeriği bakımından uygulama, coğrafik ve davranışsal bölümlenme kriterlerinden oluşmaktadır. Yapılan çalışmalar davranışsal kriterlerin bölümlenme açısından önemine vurgu yapmaktadır.

Literatürdeki çalışmalarda veri boyutunun sınırlı kalması, çalışmada kullanılan makine öğrenmesi ve kümeleme algoritmalarının istenilen performansta çalışmasına engel olmaktadır. Bu doğrultuda çalışmada kullanılan 24 milyon satır uzunluğundaki veri tabanı ve örneklem grubundaki 100.000 adet müşteri verisi, bu alanda yeni bir çalışmanın ilk adımını oluşturmaktadır. Bununla birlikte



veri tabanlı pazarlama literatüründe yapılan çalışmalar incelendiğinde, çalışmaların pazarlama süreçleri açısından sadece bir bölümü kapsadığı dikkat çekmektedir. Bu açıdan bakıldığında yapılan çalışma hem büyük veri ile ilgili adımları kapsarken, bununla birlikte müşteri kümelerinin oluşturulması, belirlenen kümelerin ileriye yönelik davranış modellerinin belirlenmesi için yapılan çoklu regresyon analizleri ve son olarak bu veriler ışığında belirlenmiş kümelere yönelik pazarlama karmalarının oluşturulması adımları ile uçtan uca tüm süreçleri kapsamaktadır.

Literatürde en çok kullanılan kümeleme algoritmalarının test edildiği bu uygulamada, DBSCAN algoritmasının, uygulamada kullanılan veri setine uygun olmaması nedeniyle kümeleme analizinde tüm verinin %91'ini bir kümeye atarak geri kalan veriler, outlier farklı bir ifade ile aykırı olarak sınıflandırılmıştır. Bu doğrultuda içerisinde demografik verinin bulunmadığı, davranışsal özelliklerin verinin anakütlesini oluşturduğu çalışmalarda K Means veya Gauss Karışım (Gaussian Mixture) gibi algoritmaların daha iyi sonuç verdiği gözlenmiştir.

Küme sayısı, bölümlendirme için önemli bir etken olarak dikkat çekmektedir. Farklı yöntemlerle birlikte, pazar hakkında bilgi sahibi olarak da belli tahminlerde bulunulabilmektedir. En yaygın kullanılan yöntemlerden biri olan Elbow yöntemi, bir fikir vermekle birlikte kesin bir sonuç ortaya koymamaktadır. Bu nedenle, farklı küme sayılarına göre kümeleme analizleri yapılmasının ve çıkan sonuçları karşılaştırılmasının en güvenilir yöntem olduğu düşünülmüştür. Bununla birlikte K means algoritmasında her farklı küme sayısı için ayrı ayrı analizlerin yapılması gerekmektedir. Bu durum, çalışmanın süresini arttırmakta ve fakat çıkan sonuçları gözlemek açısından, bir nevi oluşan kümelerin sağlamasını yapma imkanı sağlamaktadır.

Uygulama sonuçlarından ve analizlerinden yola çıkarak aşağıdaki çıkarımlar yapılmıştır:

Ürün yelpazesinin daha dar olduğu ve ana ürünün pizza olduğu bu uygulamada coğrafi özelliklerin kümeleme analizinde farklılık yaratmadığı gözlenmiştir. Bu doğrultuda ürün çeşitliliğinin fazla olduğu e-ticaret veya market sektörlerinde coğrafi özelliklerin daha fazla önem kazanacağı düşünülmüştür.

- Evlere servis siparişi ile internetten sipariş verme eğiliminin arttığı gözlemlenen bu çalışmada müşterilere, sadece hızlı olmanın yetmediği aynı zamanda tüm kanallardan sipariş verebileceğinin iletişiminin de önemli hale geldiği tespit edilmiştir.
- Artan internet siparişleriyle birlikte kredi kartı kullanımının da arttığı gözlemlenen bu çalışmada, anlaşmalı kurumlar ve bankalar aracılığı ile yapılacak çalışmaların, bu kanaldaki trafiği arttıracığı gözlenmiştir. İşletmeler açısından azalan nakit para akışının bankalar aracılığı ile toplanması, kuryelerin gün sonlarındaki hesap kapama işlemlerini, müşteri ile para üstü trafiğini azaltsa da işletmeler açısından banka komisyonlarının giderek artması, olumsuz etki olarak gözükmektedir.
- Gel al ve masaya servis tercih eden kümelerin, daha ekonomik olması sebebiyle bu kanalı tercih ettiği gözlemlenmiştir. Bununla birlikte, tüm kümelere genel çoğunluğun evlere servis siparişleri oluşturduğu; ancak hala geleneksel kanalların tercih edildiği ve bu kanallara yatırım yapılması gerektiği düşünülmüştür.
- Ürün anlamında inovasyon çalışmalarının her kümeye hitap etmese bile müşterilerin gözünde yenilikçi bir iletişim olması açısından önemli bir tutundurma faaliyeti olduğu söylenebilir. Özellikle frekansın arttığı kümelere, bu tarz ürünlerin tüketiminin çok fazla olduğu gözlenmiştir. Ancak etsiz pizza gibi daha kısıtlı tüketimi olan ürünlerin toplam satışın içerisinde payı az olsa da, menü çeşitliliği ve her kesime hitap edilebilmesi açısından işletmenin dikkate aldığı gözlenmiştir. Bununla birlikte günümüzde teknolojinin gıda ürünlerinin yapımına etkisi düşünüldüğünde, artık "Beyond Meat" gibi oluşumların bitkilerden oluşturdukları yapay et üretimiyle sadece et tüketmeyen müşteri kitlesine değil, kitlesel üretime karşı duruş sergileyen kesime de hitap eden yeni ürün geliştirdikleri bilinmektedir. Bugün birçok hızlı tüketim gıda pazarındaki büyük aktör, bu bitki bazlı etleri menülerine eklemeye başlamışlardır.

Çalışmanın önemli kısıtı, büyük verinin sadece bir firmaya ait olması ve firmanın sadece bir yıllık verilerinin kullanılarak çalışmanın gerçekleştirilmesidir. Bir diğer kısıt ise, bölümlendirme kriterlerinden sadece coğrafi ve davranışsal bölümlendirme kriterlerinin kullanılmış olmasıdır. Bunun temel sebebi ise KVKK kapsamında, veri setinde hiçbir demografik verinin bulunmamasından kaynaklanmaktadır. Bu doğrultuda yapılan analizlerde yaş, cinsiyet, eğitim durumu vb. kriterlerin kümeler üzerindeki dağılımları ve bu kriterlerdeki davranış modelleri ortaya konulamamıştır.

Kümeleme çalışmaları, veri bilimi ve analiz tekniklerinin gelişimiyle günümüzde artık her bir müşterinin kendisine özgü ve sadece kendisini kapsayan bir küme şeklinde detaylı bir alana doğru evrilmektedir. Bu doğrultuda benzerliklerden ziyade farklılıklar üzerinden yola çıkılarak yeni akademik çalışmaların yapılabileceği düşünülmüştür.

## **KAYNAKÇA**

Alpaydın, E. (2013). Yapay Öğrenme (2. Baskı). İstanbul: Boğaziçi Üniversitesi Yayınevi.

Calinski, T. Harabasz, J. (1974). A dendrite method for cluster analysis. *Communication in Statistics*. 3(1): 1-27.

Canepa, G.A. (2016). What You Need To Know About Machine Learning. Birmingham: Packt Publishing.

- Chaffey, D. (2009). E-Business and E-Commerce Management: Strategy Implementation and Practice, London: Pearson Prentice Hall.
- Cui, G. Wong, M.L. Lui, H. (2006). Machine learning for direct marketing response models: bayesian networks with evolutionary programming. *Management Science*. 52(4): 597-612.
- Davies, D.L. Bouldin, D.W. (1979). A cluser separation measure. *IEEE Transactions On Pattern Recognition and Machine Intelligence*. 1(2): 224-227.
- Dibb, S. Meadows, M. (2004). Relationship marketing and CRM: a financial services case study. *Journal of Strategic Marketing*. 12(2): 111-125.
- Durmuş, M.S. İplikçi, S. (2007). Veri kümeleme algoritmalarının performansları üzerine karşılaştırılmalı bir çalışma. Akademik Bilişim Konferansı Bildirileri. 393-400.
- Gülcan, B. (2000). Sadık müşteri yaratılabilme ve sürekli satış yapabilmenin yolu: veri tabanlı pazarlama. *Ticaret ve Turizm Eğitim Fakültesi Dergisi*. 3: 27-48.
- Harrigan, K.R. (1985). An application of clustering for strategic group analysis. *Strategic Management Journal*. 6(1): 55-73.
- Haşiloğlu, S.B. Sezgin, M. Bardakçı, A. (2008). Hizmet sektöründeki veritabanlı pazarlama araştırmalarının değerlendirilmesi. *KMU İİBF Dergisi*. 10(14): 228-240.
- Hoda, M. Jocumens G. (2003). How to implement marketing strategies using database approaches? *Journal of Database Marketing & Customer Strategy Management*. 11(2): 135-148.
- Jackson, R. Wang, P. (1996). Strategic Database Marketing. Illinois: NTC Publishing Group.
- Kantardzic, M. (2011). Data Mining: Concepts Models Methods And Algorithms. New Jersey: John Wiley & Sons Inc.
- Ketchen, D. Shook, C.L. (1996). The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal*. 17(6): 441-458.
- Kordinariya, T.M. Makwana, P.R. (2013). Review on determining number of cluster in k-means clustering. *International Journal of Advance Research in Computer Science and Management Studies*. 1(6): 90-95.
- Koslowsky, S. (1999). Reducing your risk: whats's happening in retail database marketing. *Direct Marketing*. 61(9): 40-43.
- Kotler, P. Armstrong, G. (1999). Principles of Marketing (8. Baskı). New Jersey: Prentice Hall.
- Lantz, B. (2013). Equidistance of likert-type scales and validation of inferential methods using experiments and simulations. *Electronic Journal of Business Research Methods*. 11(1): 16-28.
- MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of Fifth Berkeley Symposium on Mathematical Statistics And Probability. University of California. 1: 281-297.
- Mohri, M. Rostamzadeh, A. Talwalkar, A. (2012). Foundations of Machine Learning. London: MIT Press.
- Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. 20: 53-65.
- Sambamoorti, N. (1999). Hierarchical cluster analysis: some basics and algorithms. [http://www.crmportals.com/hierarchical\\_cluster\\_analysis.pdf](http://www.crmportals.com/hierarchical_cluster_analysis.pdf). (Erişim Tarihi: 11/12/2019)
- Turing, A. (1950). Computing machinery and intelligence: mind. *Quarterly Review of Psychology and Philosophy*. 236(59): 433-460.
- Verhoef, P.C. Spring, P.N. Hoekstra, J.C. Leeflang, P.S.H. (2003). The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands. *Decision Support Systems*. 34(4): 471-481.
- Yılmaz, Ş. Patir, S. (2011). Kümeleme analizi ve pazarlamada kullanımı. *Akademik Yaklaşımlar Dergisi*. 2(1): 91-113.

## EKLER

### Ek 1: Özelliklere Ait Açıklamalar

Özelliklere Ait Başlıklar	Anlamı	Açıklaması
%GT CustomerID	Toplam Müşteri içindeki Oranı	Küme içerisindeki müşteri sayısının toplam müşteri sayısına oranı
%GT TotalNetAmt	Ciro İçindeki Oranı	Küme içindeki müşterilerin oluşturduğu cironun, genel toplama oranı
Average of CountOrder	Frekans / Sipariş Sıklığı	Küme içindeki toplam alışveriş sayısının, küme içindeki müşteri sayısına oranı

Average of Recency	Son Sipariş Zamanı	Küme içindeki müşterilerin en son alışveriş yapma tarihinden, veri setindeki son tarihe kadar geçen sürenin küme içindeki ortalaması
Average of AvgNetAmt	Ortalama Sepet Tutarı / ATP	Küme içindeki müşterilerin ortalama sepet tutarlarının ortalaması
WebInTotal	Web Sipariş Oranı	Küme içindeki müşterilerin internetten yaptıkları alışverişlerin bütün yaptıkları alışverişe oranı
%GT TotalPizza	Toplam Pizza İçindeki Oranı	Küme içinde alınan Pizza sayısının bütün müşteriler tarafından alınan pizza sayısına oranı
SidesInTotal	Yan Ürün Oranı	Küme içindeki alınan yan ürün sayısının yapılan sipariş sayısına oranı
BuyukPizzaInTotal	Büyük Pizza	Küme içinde alınan büyük pizza sayısının toplam pizza sayısına oranı
KucukInTotal	Küçük Pizza	Küme içinde alınan küçük pizza sayısının toplam pizza sayısına oranı
ÇiftkatlıtabanInTotal	Çift Katlı Taban Oranı	Küme içinde alınan çift katlı özellikli tabanın bu küme içindeki toplam pizza alımına oranı
EtsizPizzaInTotal	Etsiz Pizza Oranı	Küme içindeki alınan etsiz pizza sayısının toplam alınan pizza sayısına oranı
AzmalzemeliPizzaInTotal	Az Malzemeli Pizza	Küme içinde alınan az malzemeli (7 malzemeli) pizza sayısının toplam pizza sayısına oranı
ÇokmalzemeliPizzanTotal	Çok Malzemeli Pizza	Küme içinde alınan çok malzemeli (13 malzemeli) pizza sayısının toplam pizza sayısına oranı
KarisikInTotal	Karışık Pizza	Küme içinde alınan ismi karışık (7 malzemeli) olan pizza sayısının toplam pizza sayısına oranı
OglenInTotal	Öğlen Sipariş Oranı	Küme içindeki 16:00'a kadar verilen sipariş sayısının toplam verilen sipariş sayısına oranı
AksamInTotal	Akşam Sipariş Oranı	Küme içindeki 16:00 ile 22:00 arasında verilen sipariş sayısının toplam verilen sipariş sayısına oranı
GecenInTotal	Gece Sipariş Oranı	Küme içindeki 22:00'den sonra verilen sipariş sayısının toplam verilen sipariş sayısına oranı
PaymentMethodType26InTotal	Kredi Kartı ile Ödeme Oranı	Kredi Kartı kullanılan sipariş sayısının toplam sipariş sayısına oranı
D0InTotal	Evlere Servis Tutarı	Evlere servis sipariş tutarı 0-15 TL arasında olan müşterilerin oranı
D1InTotal	Evlere Servis Tutarı	Evlere servis sipariş tutarı 15-30 TL arasında olan müşterilerin oranı
D2InTotal	Evlere Servis Tutarı	Evlere servis sipariş tutarı 30-45 TL arasında olan müşterilerin oranı
D3InTotal	Evlere Servis Tutarı	Evlere servis sipariş tutarı 45-60 TL arasında olan müşterilerin oranı
D4InTotal	Evlere Servis Tutarı	Evlere servis sipariş tutarı 60-75 TL arasında olan müşterilerin oranı
D5InTotal	Evlere Servis Tutarı	Evlere servis sipariş tutarı 75-90 TL arasında olan müşterilerin oranı
D6InTotal	Evlere Servis Tutarı	Evlere servis sipariş tutarı 90 TL ve üzeri olan müşterilerin oranı
C0InTotal	Gel al&Masaya Servis Tutarı	Gel al ve masaya servis sipariş tutarı 0-15 TL arasında olan müşterilerin oranı
C1InTotal	Gel al&Masaya Servis Tutarı	Gel al ve masaya servis sipariş tutarı 15-30 TL arasında olan müşterilerin oranı
C2InTotal	Gel al&Masaya Servis Tutarı	Gel al ve masaya servis sipariş tutarı 30-45 TL arasında olan müşterilerin oranı

C3InTotal	Gel al&Masaya Servis Tutarı	Gel al ve masaya servis sipariş tutarı 45-60 TL arasında olan müşterilerin oranı
C4InTotal	Gel al&Masaya Servis Tutarı	Gel al ve masaya servis sipariş tutarı 60-75 TL arasında olan müşterilerin oranı
C5InTotal	Gel al&Masaya Servis Tutarı	Gel al ve masaya servis sipariş tutarı 75-90 TL arasında olan müşterilerin oranı
C6InTotal	Gel al&Masaya Servis Tutarı	Gel al ve masaya servis sipariş tutarı 90 TL ve üzeri olan müşterilerin oranı
IstanbulInTotal	İstanbul Siparişleri	İstanbul'dan sipariş verenlerin oranı
AnkaraInTotal	Ankara Siparişleri	Ankara'dan sipariş verenlerin oranı
BursaInTotal	Bursa Siparişleri	Bursa'dan sipariş verenlerin oranı
IzmirInTotal	İzmir Siparişleri	İzmir'den sipariş verenlerin oranı
AntalyaInTotal	Antalya Siparişleri	Antalya'dan sipariş verenlerin oranı
AkdenizInTotal	Akdeniz Bölge Siparişleri	Akdeniz Bölgesinden verilen siparişlerin oranı
DoguAnadoluInTotal	Doğu Anadolu Bölge Siparişleri	Doğu Anadolu Bölgesinden verilen siparişlerin oranı
EgeInTotal	Ege Bölge Siparişleri	Ege Bölgesinden verilen siparişlerin oranı
GuneydoguInTotal	Güneydoğu Bölge Siparişleri	Güneydoğu Anadolu Bölgesinden verilen siparişlerin oranı
IcAnadoluInTotal	İç Anadolu Bölge Siparişleri	İç Anadolu Bölgesinden verilen siparişlerin oranı
KaradenizInTotal	Karadeniz Bölge Siparişleri	Karadeniz Bölgesinden verilen siparişlerin oranı
MarmaraInTotal	Marmara Bölge Siparişleri	Marmara Bölgesinden verilen siparişlerin oranı
PazartesiInTotal	Pazartesi Oranı	Pazartesi günü sipariş verenlerin oranı
SalıInTotal	Salı Oranı	Salı günü sipariş verenlerin oranı
ÇarşambaInTotal	Çarşamba Oranı	Çarşamba günü sipariş verenlerin oranı
PerşembeInTotal	Perşembe Oranı	Perşembe günü sipariş verenlerin oranı
CumaInTotal	Cuma Oranı	Cuma günü sipariş verenlerin oranı
CumartesiInTotal	Cumartesi Oranı	Cumartesi günü sipariş verenlerin oranı
PazarInTotal	Pazar Oranı	Pazar günü sipariş verenlerin oranı

## Ek 2: Tüm Özelliklerin Veri Setinde Bulunduğu Dörtlü Küme Sonuçları

4 Kmeans	A	B	X	C
%GT CustomerID	85,00%	13,24%	0,01%	1,74%
%GT TotalNetAmt	50,70%	35,32%	0,55%	13,43%
Average of CountOrder	1,56	7,12	157,86	20,95
Average of Recency	157,76	59,85	8,86	27,23
Average of AvgNetAmt	35,59	35,27	52,61	34,35
WebInTotal	49,42%	65,27%	76,74%	67,02%
%GT TotalPizza	51,30%	35,82%	0,55%	12,32%
SidesInTotal	41,67%	49,01%	15,48%	62,03%
BuyukPizzalnTotal	19,17%	20,08%	66,17%	19,06%
KucukInTotal	15,09%	16,14%	26,56%	18,92%
ÇiftkatlitabanInTotal	1,43%	2,65%	0,00%	5,12%

EtsizPizzalnTotal	4,49%	5,53%	1,48%	7,59%
AzmalzemeliPizzalnTotal	24,00%	24,52%	75,85%	22,63%
ÇokmalzemelipizzanTotal	21,08%	19,79%	1,91%	17,57%
KarisikInTotal	17,01%	11,39%	0,18%	10,87%
OgleInTotal	23,32%	24,08%	35,29%	25,71%
AksamInTotal	58,70%	59,20%	50,05%	55,29%
GecenInTotal	17,98%	16,71%	14,66%	18,99%
PaymentMethodType26InTotal	57,80%	49,79%	88,14%	50,62%
D0InTotal	1,95%	2,63%	2,26%	3,52%
D1InTotal	36,52%	38,38%	1,00%	42,25%
D2InTotal	32,46%	29,91%	0,90%	28,08%
D3InTotal	9,74%	9,54%	0,09%	8,91%
D4InTotal	4,44%	4,53%	0,00%	4,43%
D5InTotal	1,16%	1,23%	0,00%	1,19%
D6InTotal	0,07%	0,07%	6,70%	0,21%
C0InTotal	2,59%	3,04%	27,06%	2,67%
C1InTotal	6,08%	5,98%	44,34%	5,11%
C2InTotal	2,75%	2,57%	13,30%	1,90%
C3InTotal	0,95%	0,97%	2,99%	0,68%
C4InTotal	0,32%	0,31%	1,18%	0,24%
C5InTotal	0,10%	0,08%	0,09%	0,08%
C6InTotal	0,07%	0,07%	6,70%	0,21%
IstanbulInTotal	37,95%	39,35%	55,48%	42,33%
AnkaraInTotal	10,07%	10,35%	11,22%	8,51%
BursaInTotal	4,94%	5,17%	22,17%	6,22%
IzmirInTotal	7,39%	7,51%	10,41%	8,08%
AntalyaInTotal	3,14%	4,05%	0,18%	3,21%
AkdenizInTotal	8,81%	8,92%	0,18%	6,40%
DoguAnadolunTotal	1,68%	1,43%	0,00%	1,37%
EgeInTotal	12,88%	13,19%	10,41%	13,42%
GuneydogunTotal	2,97%	2,54%	0,00%	2,40%
IcAnadolunTotal	14,50%	14,40%	11,22%	11,62%
KaradenizInTotal	6,53%	5,28%	0,00%	6,40%
MarmaraInTotal	52,64%	54,24%	78,19%	58,39%
PazartesInTotal	13,79%	14,05%	17,19%	14,83%
SaliInTotal	12,21%	12,51%	13,85%	13,29%
CarsambaInTotal	13,54%	13,97%	12,13%	14,25%
PersembenInTotal	14,87%	15,40%	17,83%	15,57%
CumalnTotal	14,90%	15,34%	13,21%	15,23%
CumartesInTotal	13,62%	12,71%	13,57%	12,08%
PazarInTotal	17,07%	16,02%	12,22%	14,75%

## Ek 3: Tüm Özelliklerin Veri Setinde Bulunduğu Dörtlü Küme Sonuçları

2 Kmeans	A1	A2
%GT CustomerID	63,99%	36,01%
%GT TotalNetAmt	75,21%	24,79%
Average of CountOrder	2,2	1,41
Average of Recency	36,87	33,30
Average of AvgNetAmt	73,91	284,68
WebInTotal	54,76%	47,27%
%GT TotalPizza	74,23%	25,77%
SidesInTotal	43,06%	43,62%
BuyukPizzalnTotal	19,46%	18,88%
KucukInTotal	16,02%	12,98%
ÇiftkatlıtabanInTotal	1,59%	1,52%
EtsizPizzalnTotal	4,75%	4,51%
AzmalzemeliPizzalnTotal	24,53%	23,26%
ÇokmalzemelipizzanTotal	20,00%	23,40%
KarisikInTotal	15,12%	17,67%
OgleInTotal	23,39%	23,27%
AksamInTotal	59,12%	59,04%
GeceInTotal	17,48%	17,69%
PaymentMethodType26InTotal	55,35%	58,30%
D0InTotal	1,71%	3,40%
D1InTotal	35,05%	41,62%
D2InTotal	32,37%	30,81%
D3InTotal	10,28%	8,23%
D4InTotal	5,02%	3,27%
D5InTotal	1,27%	0,85%
D6InTotal	0,09%	0,05%
C0InTotal	2,31%	3,19%
C1InTotal	6,55%	4,70%
C2InTotal	2,96%	2,22%
C3InTotal	1,02%	0,72%
C4InTotal	0,36%	0,30%
C5InTotal	0,09%	0,09%
C6InTotal	0,09%	0,05%
IstanbulInTotal	37,94%	38,84%
AnkaraInTotal	10,16%	10,18%
BursaInTotal	5,12%	4,15%
IzmirInTotal	7,20%	7,09%
AntalyaInTotal	3,56%	3,25%
AkdenizInTotal	8,48%	8,95%
DoguAnadoluluInTotal	1,71%	1,68%
EgeInTotal	13,16%	12,50%
GuneydoguluInTotal	2,87%	3,04%

IcAnadolulnTotal	14,42%	15,08%
KaradenizInTotal	6,41%	6,39%
MarmaraInTotal	52,96%	52,36%
PazartesiInTotal	13,63%	14,23%
SaliInTotal	12,29%	11,88%
CarsambaInTotal	13,50%	13,33%
PersembelInTotal	14,75%	15,70%
CumaInTotal	15,04%	14,60%
CumartesiInTotal	13,30%	14,44%
PazarInTotal	17,49%	15,82%