T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

OTEL REZERVASYON VERİLERİNE GÖRE SATIŞ İÇİN MÜŞTERİ PROFİLİ BELİRLEME

Yüksek Lisans Tezi

SALİH CANTEKİN

İSTANBUL, 2020

T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİ TEKNOLOJİLERİ

OTEL REZERVASYON VERİLERİNE GÖRE SATIŞ İÇİN MÜŞTERİ PROFİLİ BELİRLEME

Yüksek Lisans Tezi

SALİH CANTEKİN

Tez Danışmanı: DR. ÖĞR. ÜYESİ BETÜL ERDOĞDU

İSTANBUL, 2020

T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİ TEKNOLOJİLERİ

Tezin Adı: Otel Rezervasyon Verilerine Göre Satış İçin Müşteri Profili Belirleme
Öğrencinin Adı Soyadı: Salih CANTEKİN
Tez Savunma Tarihi: 22.06.2020

Bu tezin Yüksek Lisans tezi olarak gerekli şartları yerine getirmiş olduğu Fen Bilimleri Enstitüsü tarafından onaylanmıştır.

Doç. Dr. İsmail Burak KÜNTAY
Enstitü Müdürü

Bu tezin Yüksek Lisans tezi olarak gerekli şartları yerine getirmiş olduğunu onaylarım.

Bu Tez tarafımızca okunmuş, nitelik ve içerik açısından bir Yüksek Lisans tezi olarak yeterli görülmüş ve kabul edilmiştir.

Jüri Üyeleri	İmzalar
Tez Danışmanı Dr. Öğr. Üyesi Betül ERDOĞAN	----------
Üye Dr. Öğr. Üyesi Zeynep Turgut	----------
Üye Dr. Öğr. Üyesi Özge Yücel Kasap	----------

ÖZET

OTEL REZERVASYON VERİLERİNE GÖRE SATIŞ İÇİN MÜŞTERİ PROFİLİ BELİRLEME

Salih CANTEKİN

Bilgi Teknolojileri Yüksek Lisans Programı

Tez Danışmanı: Dr. Öğr. Üyesi Betül ERDOĞDU

Haziran 2020, 40 sayfa

Firmalardaki çok büyük ölçekli olarak tanımlanabilen ve yüzbinlerce veriye sahip yazılım sistemleri üzerinden, ihtiyacı karşılaması için değerli verinin elde edilmesi işlemi sürecine veri madenciliği denilmektedir. Bu sayede ileriye yönelik tahminlerde bulunabilmeyi kolaylaştırmak için verilerin birbiri ile olan ilişkilerinin ortaya konması kolaylaşmıştır. Veri madenciliği ile milyarlarca veri üzerinde çalışabiliyoruz. Bu madenciliğin temel amacı, kurumlarda bulunan karar destek sistemleri için değerli olacak bilgiyi belirli yöntemler, algoritmalar ve işlem süreçleri uygulayarak ortaya çıkarmaktır.

Bu çalışma, elde bulunan büyük ölçekteki müşteri verileri üzerinden, veri madenciliği metotları ile müşteri profili çıkarmayı amaçlamaktadır. Bu çalışmada, turizm sektöründe halen aktif olarak faaliyet gösteren bir turizm firmasına ait, kişisel verilerin korunması kanunu gereğince isimsiz bir veri kümesi kullanılmıştır. Bu veri üzerinde çeşitli kümeleme algoritmaları ve açıklayıcı veri analiz yöntemi kullanılarak müşteriler niteliklerine göre gruplara ayrıştırılmış.

Anahtar Kelimeler: Müşteri Profili Belirleme, Veri Madenciliği

ABSTRACT

CUSTOMER PROFILE DETERMINATION BY HOTEL RESERVATION DATA

Salih CANTEKİN

Information Technology

Thesis Supervisor: Assist. Prof. Dr. Betül ERDOĞDU

June 2020, 40 pages

The process of obtaining valuable data to meet the need through software systems, which can be defined on a very large scale and with hundreds of thousands of data, is called data mining. In this way, it is easier to reveal the relationships between the data in order to make it easier to make forward predictions. With data mining, we can work on billions of data. The main purpose of this mining is to reveal the information that will be valuable for the decision support systems in the institutions by applying certain methods, algorithms and processing processes.

This study aims to create a customer profile with data mining methods on the large amount of customer data available. In this study, an anonymous data set was used in accordance with the law of protection of personal data, belonging to a tourism company, which is still active in the tourism sector. The customers were divided into groups according to their qualifications using various clustering algorithms and descriptive data analysis method.

Keywords: Customer Profile Determination, Data Mining

İÇİNDEKİLER

ŞEKİLLER

1. GİRİŞ

Çok uzun bir süredir ortalarda dolaşan ve hemen hemen herkesin de katıldığı fikir; Bilgi Güçtür! Bilgi üretebilen toplumlar, şirketler ve bireyler daha hızlı gelişir, refah düzeylerini karlılıkları ve gelirleri ile birlikte artırırlar. İçerisinde bulunduğumuz bilgi çağında farkındalık ve değer yaratmanın en önemli yolu fizikler varlık göstermekten ziyade, bilgi kaynaklarının etkin kullanılmasından geçiyor. Bu sebeple bilginin yönetimi için birçok araç ve yöntem geliştiriliyor. Günümüzde, gelişmekte olan bilgi teknolojileri sayesinde her gün daha çok sayısal anlamda veri toplanıyor, saklanıyor ve daha da önemlisi kullanılıyor[1].

Firmalar geleceğe dönük kararlar almak istediklerinde, elektronik ortamdaki verileri analiz etmek isterler. Veri madenciliği sayesinde, firmaların bilgisayar ortamlarında yığınlar halinde tuttukları veriler anlam kazanır.

Veri ham haliyle duruyorsa ve işlenmiş ise, değersizdir. Yığınlar halindeki veriler işlenerek değerli veri haline getirilerek faydalı bilgiye dönüştürülebilir. En temelde veriler özelliklerine göre önceden tanımlanmış olan sınıflara ayrılır.

Yığın halindeki verilere anlam kazandırma işi veri madenciliği sürecidir. Veri madenciliği bunu, kayıt altına alınmış olan yığın halindeki verileri gerekli işlemlerden geçirerek yapar ve sonucunda bilgiye dönüştürür. Bu şekilde bilgiye dönüşen veriler, satış ve pazarlama hamlelerinde önemli rollere sahip olurlar.

Örneğin; marketler kendilerinden alışveriş yapan herkesin her türlü bilgisini saklamak ister. Bunların bir kısmını, internet üzerinden veya market içerisindeki anketler sonucu elde ederler. Daha sonrasında o müşterilerin yaptıkları her türlü alışverişi saklarlar. Ancak kayıt altına alınan ve anlamsız görünen bu veriler veri madenciliği süreci sayesinde analiz edilip, kategorize edilip müşteriler alışveriş yapma eğilimlerine göre ayrılır. Daha sonra

ise müşterilerin hangi ürün, hizmet ve satın alımlarla ilgilenebilecekleri tahminlenerek o yönde kampanya, satış veya pazarlama stratejilerini belirlerler. Yada çevrimiçi alışveriş sitelerinde yapmış olduğunuz eski alışveriş kayıtları incelenerek size ilgilenebileceğiniz yeni ürünler için teklif verebilirler[2].

Bir zamanların en büyük cep telefonu üreticisi olan Nokia'nın çöküşündeki hikâye de bu konudaki en çarpıcı örneklerden birisidir. Teknoloji sektöründe etnograf olarak görev yapan Tricia Wang 2017 yılında Çin'deki alt gelir grubundaki insanların piyasaya yeni girmiş olan akıllı telefonlara olan ilgisini gözlemleyerek Nokia şirketine bu konuda alınması gereken önlemleri de içeren bir rapor sunmuş ancak bu rapor firma tarafından pek önemsenmemiştir. Sonucunda ise kısa bir süre içerisinde firmanın satışları azalır ve piyasadan silinme noktasına gelir[3].

M. Canbay ve N. Duru 2007 yılında veri madenciliğini kullanarak deprem verilerinin analiz edilmesi üzerine bir çalışma gerçekleştirmiştir. Yapılan bu çalışmada, deprem verilerini kullanarak seçilmiş bölgelere ait sismik tehlikelerin başka bir söyleyişle depremlerin gerçekleşme olasılığını veri madenciliği ile ele alarak inceleme yapılmıştır. Çalışmanın sonuçları jeofizik sonuçlar ile birlikte korelasyon yöntemi ile analiz edilmiş ve doğruluk payları da artırılmıştır. Gelecekteki her 10 yıl için sismik tehlike değeri artış göstermiş ve bu şekilde devam etmiştir. Örneğin; 6 şiddetindeki bir depremin gerçekleşme olasılığı 10 yıl içinde yüzde 27 değerine sahipken, 30 yıl içerisinde önce yüzde 60 seviyesine daha sonra 60 yıl içerisinde de yüzde 80'ler bulmaktadır. Bu şekilde ortaya çıkmış veriler, çalışma bölgesinde daha önce yapılmış olan çalışmalarla karşılaştırıldığında sonuçlar birbiri ile uyum göstermektedir. Kullanılan bu teknik, depremlerin gerçekleşme tahminlemesi için kullanılan tekniklerden sadece birisidir. Çalışmaların konusu itibariyle ortaya çıkmıştır ki, çalışma yapılan bölgelerdeki tektonik veriler hiç incelenmeden dahi bu konuda olumlu sonuçlar ortaya çıkarmak mümkündür. Ayrıca bu çalışma, küçük bölgelerdeki sonuçların büyük bölgelere göre daha verimli ve

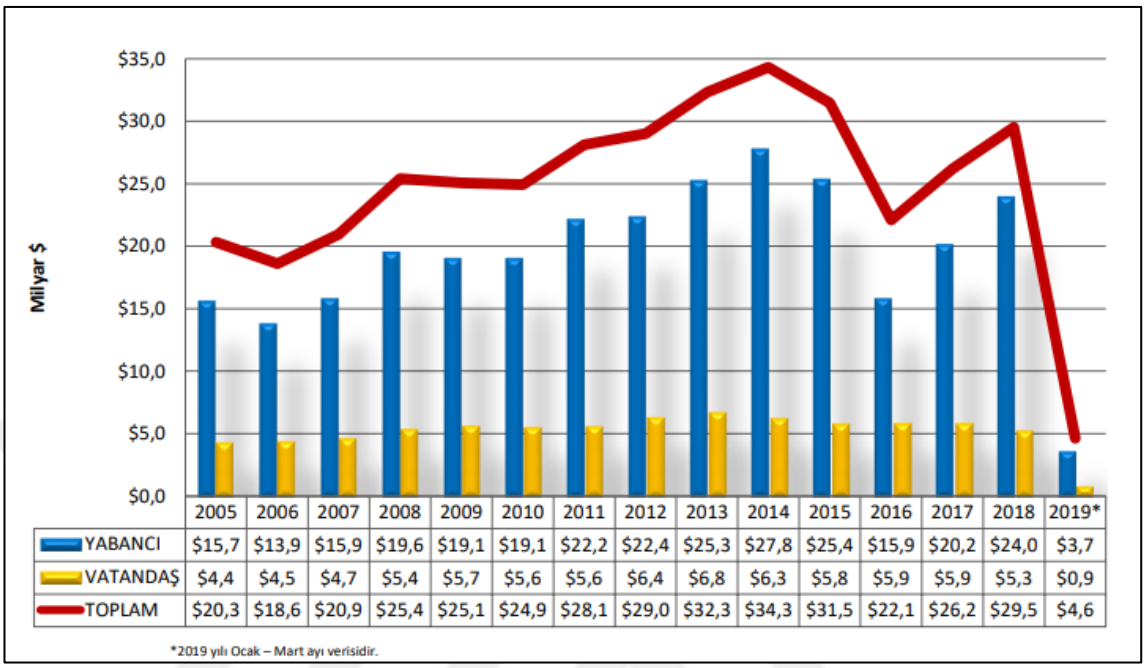
doğruluğu yüksek olduğunu da göstermiştir. Uygulama, dünya ölçeğindeki her nokta için aynı analizi yapacak şekilde geliştirilerek ihtiyaç duyulması halinde program üzerinde eklemeler yapılarak da aynı çalışmaların yapılması için tasarlanmıştır[4].

Turizm sektörü, Türkiye ekonomisinin pasta paylarına bakıldığında en büyük paya sahip lokomotiflerinden biridir. Turizmin esas geliri ise turistin kendisidir. Bu sebeple turizmde kalkınma hareketleri planlanırken düşünülmesi gereken asıl konu turist sayısının artırılması olmalıdır. Ülkeye giriş yapan turistlerin beklentileri ve amaçlar genel anlamı ile bilinmektedir. Turistleri tanımak, ileriki yıllarda kendilerine daha iyi hizmet ve ürün sunabilmek ve pazarlayabilmek için önemlidir. Bunun için de turizm konusunda ayrıntılı istatistiklere ihtiyaç duyulmaktadır[5].

Türofed'in (Türkiye Otelciler Federasyonu) 2019 Mart ayındaki bültenine göre, Türkiye'nin turizm gelirleri 2018 yılında, bir önceki yıla göre yüzde 12.5 oranında artış göstererek 29.5 milyar Amerikan Doları olarak gerçekleşmiştir. Öte yandan yurtdışında yaşayan vatandaşlarımızdan elde etmiş olduğumuz gelir de 5.3 milyar Amerikan Doları olarak gerçekleşmiş ve önceki yıl verileri ile karşılaştırıldığında yüzde 10.1 oranında gerileme gözlenmiştir. 2019 Ocak-Mart ayları temel alındığında ise 4.6 milyar Amerikan Doları turizm geliri elde edilmiştir[6].

Şekil 1.1'de 2005-2019 yıllarına ait gelir grafiği rakamları verilmiştir.

Şekil 1.1: Yabancı ziyaretçi ve yurtdışında ikamet eden vatandaş ziyaretçi turizm gelirlerinin yıllara göre dağılımı



	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019*
YABANCI	$15,7	$13,9	$15,9	$19,6	$19,1	$19,1	$22,2	$22,4	$25,3	$27,8	$25,4	$15,9	$20,2	$24,0	$3,7
VATANDAŞ	$4,4	$4,5	$4,7	$5,4	$5,7	$5,6	$5,6	$6,4	$6,8	$6,3	$5,8	$5,9	$5,9	$5,3	$0,9
TOPLAM	$20,3	$18,6	$20,9	$25,4	$25,1	$24,9	$28,1	$29,0	$32,3	$34,3	$31,5	$22,1	$26,2	$29,5	$4,6

*2019 yılı Ocak – Mart ayı verisidir.

Bütün turizm firmaları bu pazarda kendilerine düşen pasta payını büyütmek için rekabet halindedirler. Bu rekabet ortamında üst sıraları hedefleyebilmek için sahip oldukları verileri en iyi şekilde yönetmeli ve işleyebilmelidirler. Gelişen teknoloji, günümüzde veriyi çok efektif bir şekilde işlemeye olanak sağlamaktadır. Bu teknolojik gelişmeler ışığında firmalar, kendileri için en doğru müşteri profillerini belirleyebilirler ise, hem reklam için harcadıkları maliyetlerini düşürebilirler hem de kendileri için uygun kitleyle doğrudan iletişime geçerek hızlı geri dönüşler sağlayabilirler. Veri madenciliği yöntemleri ile ürününüzün çoğunlukla genç nüfus tarafından tercih edildiğini tespit ettiğinizi varsayalım. Gençlere hitap eden bu ürününüzün reklamları için neden GSM operatörlerine bütün yaş grupları için para ödeyesiniz ki? Veri madenciliği ile mevcut müşteriler daha iyi tanınarak firmaların müşteri ilişkileri departmanlarında düzenleme ve geliştirmeler yapılabilir. Müşterinin iyi tanınması, kişiselleştirilmiş ürün ve hizmetlerin tanıtılmasını mümkün kılacaktır. Bu sayede firmaların "müşteri gibi düşünme" empatisi de geliştirilebilir. Bu çalışmadaki amacımız, sahip olduğumuz turizm sektörüne ait bu

büyük ölçekli veriyi, veri madenciliği metotları ile analiz ederek müşteri profillerini belirlemek. Veri madenciliği işlemleri sonrasında oluşan modeller, bize bu kümeler içerisindeki müşterilerin ortak özellikleri hakkında bilgiler verecek ve onların gelecekte yapacakları tercihleri tahmin etme konusunda bizlere olanak sağlayacaktır. Çalışmada turizm sektöründe halen aktif rol alan bir firmanın verileri kullanılmıştır. Veriler en yaygın kümeleme tekniklerinden biri olan *k-means* tekniği ile farklı kümelere ayrıştırılmıştır. Bu verilerin nasıl toplandığı, hangi boyutlara göre analiz edildiği ve elde edilen bulguları ilerleyen bölümlerde anlatılacaktır.

1.1 VERİ MADENCİLİĞİ VE BİLGİ KEŞFİ

Dünya üzerinde otomasyon verileri, tıp alanındaki veriler ve bunun gibi birçok alanda elde edilmiş veri miktarları günümüzde öngörülemez şekilde hızlı artmaktadır. Böylece bu verilerin toplanması ve saklanması gibi problemler ortaya çıkmıştır. Dosya sistemleri ve veri tabanlarındaki birtakım gelişmeler bu problemlere çözüm olmaya çalışmaktadır. Veri tabanlarındaki büyük çaptaki gelişmeler ve bilgisayar donanım ürünlerinin birçok insan ve kurum tarafından karşılanacak bilecek kadar ucuzlamış olması problemlere çözüm bulma noktasında işleri kolaylaştırmaktadır. Toplanan verilerin boyutlarının çok büyük olmasına karşın bu verilerin yalnızca küçük bir kısmı kullanılıyor. Boyutların bu kadar büyük olması, herhangi bir veri işleme ve analiz aracı kullanmadan bu verilerin analiz edilerek karar destek aşamalarında kullanılmasını imkânsız kılmıştır.

Çoğu zaman veri tabanlarında tutulan ve iyi kullanılmayan veriler insanlar için de çözülmesi gereken bir problem haline gelebilmektedir. Daha iyi çözümleme tekniklerine olan gereksinimin artma sebeplerinden birisi de toplanan verinin miktarı ve verilerdeki karmaşıklık miktarının artmasıdır. İşte tam da bu noktada Veri Tabanlarında Veri Madenciliği ve Bilgi Keşfi kavramları ortaya çıkmaktadır. Keşfedilmiş bilginin, organizasyonlarda karar verme süreçlerinin geliştirilmesinde kullanılması için, veri tabanlarındaki bilgi keşfi sürecinin başarılı olması gerekir. Bir firmanın veri tabanında

bulunan geçmiş dönem satış bilgileri, “herhangi bir müşteri grubunun sunmuş olduğunuz hangi hizmetlere ilgi duyabileceği” şeklinde bir bilgi içerebilir. Bunun analizinin yapılması firmanın satış rakamlarının artmasına sebep olabilir. Bilgi Keşfi ve Veri Madenciliği kavramları her ne kadar benzer görünüyor olursa olsun aralarında belli farklar mevcuttur. Bilgi Keşfi, yığın halindeki verilerden anlamlı bilgilerine elde edilmesi için gerekli olan süreçlerin tümünü tanımlar. Ancak veri madenciliği, Bilgi Keşfi sürecindeki önemli adımlardan sadece bir tanesini temsil eder. Ayrıca, veri madenciliği, bilgi keşfi sürecinde model belirlenmesi ve değerlendirmesi aşamalarından meydana gelmiş önemli bir kesimdir. Farklı farklı veri kaynaklarından verilerin toplanması işlemi sürecin başlangıç adımı iken, toplanmış olan bu verinin analiz edilebilmesi için en uygun ve faydalı hale getirilmesi sonraki adımı oluşturmaktadır. Veri ambarlarına sahip kuruluşlarda, gerekli olan veri deposu olarak adlandırılmış olan daha küçük çaptaki veri ambarların aktarılarak doğrudan Veri Madenciliği süreci işlemleri de başlatılabilmektedir[7].

1.2 VERİ MADENCİLİĞİ KAVRAMI

Aşağıda bazılarına da yer verilmiş veri madenciliği için farklı tanımlamalar yapılmıştır.

Shapiro & Fayyad’a göre, veri içerisinden anlamlı örüntüler ve ilişkiler elde etme sürecine bilgi keşfi, veri madenciliği veya veri arkeolojisi gibi isimler de verilmiştir. Bu tanımı daha çok veri analistleri, yönetim bilişim sistemleri kullanıcıları ve istatistikçiler kullanmaktadır. Veri tabanlarındaki bilgi keşfi tanımlaması ilk kez 1989 yılındaki bir atölye çalışması sırasında, bilginin son ürün olduğunu belirtmek için kullanılmıştır[8].

Özmen veri madenciliğini, oldukça hızlı toplanan ve büyük miktardaki verilerin çeşitli analizlerin uygulanması sonucu olarak daha anlamlı bilgilere dönüştürülmesi sırasında devreye giren bir süreç olarak görmektedir[9].

Intera Systems şirketinin resmi web sitesinde veri madenciliği kavramı, arşivlenmiş veriler üzerinden doğru analizlerin yapılması sonucunda önceden bilinmeyen ancak değeri yüksek ve anlaşılabilir sonuçların çıkarılması süreci olarak tanımlanmaktadır. Elde edilen sonuçlar tahmin yürütme, sınıflandırma veya veriler arasındaki gerekli benzerliklerin bulunması amacı doğrultusunda değerlendirilir ve özellikle de karar destek sistemlerinde kullanılır[10].

Linoff ve Berry'e göre, veri madenciliği büyük miktarlardaki verilerin kurallar ve anlamlı olan örüntülerinin çıkarılması amacı ile analiz edilmesi işlemidir. Veri madenciliği tahmin, görselleştirme, kümeleme, birliktelik kuralları ve sınıflandırma gibi yöntemlerden yararlanmaktadır[11].

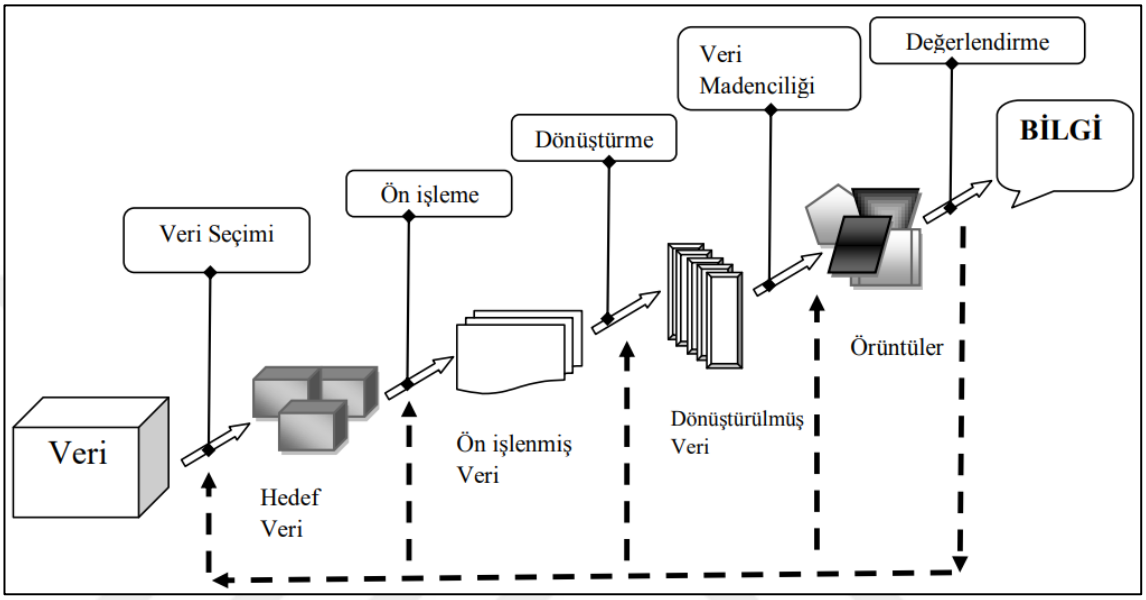
Zantinge ve Adrians'a göre veri tabanları içerisinde zenginleştirilmiş bilgilere sahip olan birçok firma, bu bilgiyi yönetmenin hayli zor olması nedeniyle bilgisayar gibi teknolojik aletleri kullanmaktadır. Özellikle de bilgisayarların gerekli araçlar ile birlikte kullanılarak veri içerisinden doğru ve anlamlı bilgilerin çıkarılması veri madenciliği olarak tanımlanır[12].

Alpaydın'a göre veri madenciliği, veri ambarlarından saklanan ve büyük miktarda olan veriler içinden, gelecek hakkında tahmin yapmamızı sağlayacak kuralların ve bağlantıların bilgisayar programlarının kullanılarak aranması işlemidir. Bu verileri analiz ederek, bir ürün veya hizmet için sonraki ayın satışları hakkında tahminleme yapılabilir, müşteriler aldıkları hizmetlere göre sınıflandırılabilir veya yeni ürünler için potansiyel müşteriler belirlenebilir[13].

Büchner ve Anand'ın veri madenciliğini için kullandığı tanım, çok büyük veri kümelerinden önemi olan ancak daha önceden bilinmeyen, anlaşılır ve yararlı örüntülerin keşfedilmesidir[14].

Jothishankar, veri madenciliğini bilgisayar işlemcilerine nasıl soracağımızı bilemediğimiz sorularımızın cevaplarını istemenin bir yolu[15].

Şekil 1.2: Veri Madenciliği Süreci



1.3 TEKNOLOJİNİN TURİZM İŞLETMELERİ ÜZERİNDEKİ ETKİSİ

Teknolojide meydana gelen gelişmeler sadece seyahat etmeyi seven insanların tercihlerinde değişimlere sebep olmadı, aynı zamanda müşterilerini tanımak ve onlara taleplerini karşılayacak şekilde seyahat seçenekleri sunmak isteyen firmaların da faaliyetlerinde değişikliğe gitmesine sebep oldu. Günümüzde artık birçok firmanın kendi sektöründe tutunabilmesi, kullandığı teknolojilere hakim olması ve bu teknolojileri kullanmaları ile doğru orantılıdır. Firmalar artık müşterilerinin medeni durumu, yaşı, gelir düzeyi, çocuklarının olup olmadığı, eğitim durumu gibi demografik verilerden boş vakitlerinde neler yapmaktan hoşlandıklarına, nerelere seyahat etmeyi sevdiklerinden kaç yaşında evlendiği bilgisine kadar çok çeşitli veriyi saklayıp işleyerek, müşterisini doğru tanıyıp doğru pazarlama teknikleri kullanarak piyasadaki rakiplerinin önüne geçme amacı gütmektedirler.

Turizm alanında faaliyet gösteren firmaların müşterilerine sunmuş olduğu çeşitli hizmetler satış adımlarında denetlenebilir veya kontrol edilebilir ürünler değildir. Bu ürünler çok uzaklardaki hatta iklimi bile farklı olan bir coğrafyadaki bir otel, aynı şehirdeki bir kültür turu, denizleri ya da okyanusları sevenler için bir Cruise Gemi turu veyahut bir ibadet ziyareti olabilir. Bu ürünleri müşterilerine satmak isteyen bir turizm firması elindeki tüm ürünlerin müşterileri tarafından daha iyi tanınmış olması için teknolojiyi kullanmalıdır. Gerekli tüm bilgi altyapısı bu teknolojik enstrümanlarla sağlanarak müşterilerinin beklentilerini karşılamalıdır.

Günümüzde turizm firmaları, rezervasyon sistemleri, mobil uygulamalar ve web siteleri gibi bilgi yönetim sistemlerini kullanmaya ve dünya çapında sistemlerin içerisinde yer almaya başlamıştır. Bulunduğu yeri korumak isteyen firmalar interneti ve internet sistemlerini efektif olarak kullanma zorunluluğuna girmiştir. Rekabet stratejilerini günden güne geliştirmek ve pazardaki gereksinimleri en iyi şekilde yorumlayabilmek artık teknolojinin efektif kullanımı ile doğru orantılı hale gelmiştir. Günümüzde bilgi teknolojilerinin ve kullanımının ucuzlaması, firmaların uluslararası düzeydeki değişimlerini ve dolayısı ile globalleşmesini hızlandırmıştır. Böylelikle cep telefonları, bilgisayarlar, tabletler gibi donanımları kullanabilen insanların yeni iş olanaklarına sahip olmasının da önü firmaların da aynı teknoloji ve donanımları kullanması sayesinde açılmıştır.

Firmaların ekonomik anlamdaki etkinliklerinin artması, internet kanalları ile yapılan e-ticaret, yeni pazarlama teknikleri veya dijital pazarlama etkinliklerinin artması ile sağlanabilir durumdadır. Turizmin gelişmekte olan ülkelerin ekonomik olarak kalkınmasındaki rolü yadsınamaz bir gerçek haline gelmiştir. Turizm firmalarının dünyadaki seyahat ve turizm endüstrisindeki büyük değişimlere ayak uydurabilmesinin tek yolu, meydana gelen bu gelişim ve değişimlere mümkün olunan en hızlı şekilde uyum sağlamaktan geçiyor. Yine aynı şekilde global turizm pastasından en büyük payı almak isteyen ülkeler, kendi turizm politikalarını oluştururken, tüm gelişim ve değişimleri en doğru şekilde yorumlayarak politikalarını bu yöne uygun olarak oluşturmalıdırlar. Dünya

da turizm endüstrisi tarafından yine aynı endüstriye gerekli olan tüm bilgileri sağlamak için koşullar gözetilerek oluşturulmuş global dağıtım sistemleri vardır. Bu sistemler otel rezervasyonu, gezi veya kültür turları gibi turizm firmalarının müşterilerine sunmuş olduğu hizmetlerin bir araya getirilmesini sağlayan sistemlerdir. Bu sistemler en başlarda sadece seyahat acentaları tarafından kullanılıyordu ancak internet hizmetlerinin yaygınlaşması ve internet hızının artması ile birlikte tüm bu sistemler müşterilerin de kullanabileceği dijital platformlar haline gelmiştir.

Amadeus, Worldspan, Galileo ve Sabre bu global dağıtım sistemleri arasındaki en büyük ve en önemlileri olarak gösterilir. Turizm firmaları global dağıtım sistemlerinden aldıkları teknik destek ve veriler ile birlikte kendi müşterilerine en iyi seçenekleri sunabilmektedir. İnternet üzerinden de kolayca erişilebilen bu sistemler turizm firmalarının müşterilerine sunduğu tüm hizmetleri bir araya getirerek, müşterilerine çeşitli alternatifler arasından en uygun olanın bulunması konusunda yardımcı olmaktadır.

Sürekli gelişim ve değişimleri 90'lı yılların başından itibaren devam ettiren bu sistemler firmalara sebep oldukları maliyetler ile karşılaştırıldığında diğer birçok yönteme göre daha uygulanabilir olması sebebiyle daha fazla fayda sağlamaktadırlar. Tüm dünya ile birlikte turizm firmalarının ve teknolojinin de gelişmeye devam etmesi, Amadeus, Sabre gibi firmalarında kendini geliştirmeye devam ederek sunulan hizmetlerin ve hizmet kalitesinin artmasına katkıda bulunacaktır[16].

2. LİTERATÜR TARAMASI

2020 yılındaki güncel rakamlar, yaklaşık olarak 80 milyonluk nüfusa sahip bir ülkede yaşadığımızı gösteriyor. Bu nüfusu oluşturan her birey, fiziksel, kişisel, duygusal ve daha birçok özelliklerine göre birbirinden ayrılıyor. Bu kadar farkın olduğu koşullar incelendiğinde, her birimizin tamamen aynı davranışları sergiliyor olması ne kadar mümkün olabilir dersiniz? Çok mümkün görünmüyor. Satış ve Pazarlama stratejilerinin ana konusunun insan faktörü olduğunu düşünürseniz, insanların da göstermiş olduğu davranışsal farklılıkların bu stratejileri doğrudan etkilediğini söyleyebiliriz. Tam da bu noktada bahsi geçen stratejilerin etkinliğinin artması ve daha efektif olması için müşterileri profillerine göre kategorize etmek büyük önem arz ediyor. Satış ve Pazarlama için müşterilerinizi hedeflemenin yolu; nasıl davrandıklarını, nerede olduklarını, ne zaman ve nasıl hareket ettiklerini ve en önemlisi harekete geçip geçmediklerini bilmekten geçiyor.

2.1 PEKİ NEDİR BU MÜŞTERİ PROFİLİ?

En basit ifadesiyle müşteri profili; müşterilerinin tipinin açıklamasıdır. Genel olarak müşteri davranışlarını, özelliklerini ve bu özellikleri anlamak, anladıklarınızı satış ve pazarlama faaliyetleriniz doğrultusunda kullanmak amacını kapsar.

Müşteri profillerinin oluşturulması yeni bir kavram değil elbette. Ancak dijital çağın ve teknolojinin de yardımıyla müşterileriniz hakkında bilgi toplama şeklimizin değişmesi sebebiyle profillerin oluşturulması konusuna bakış açımız değişti. Potansiyel müşterileriniz hakkındaki veriler deyim yerindeyse daha parmaklarının ucundayken, bu veriyi en etkili şekilde nasıl kullanacağınızı planlamak ve bunlar ile ilgili gerekli aksiyonları almak hayati öneme sahip.

Yavuz Odabaşı ve Gülfidan Barış (2011) tarafından kaleme alınan kitapta, tüketici davranışları 7 ana kategoride değerlendirmeye alınmış. Bu davranışların güdülenmiş olduğu, dinamik bir süreç olduğu, çeşitli faaliyetlerden oluştuğu, karmaşık ve zaman açısından farklılıklar gösterdiğini, farklı rollerle ilgilendiğini, çevre faktörlerinden etkilendiğini ve farklı kişiler için farklılıklar gösterdiğini belirtmişlerdir[17].

Vahaplar ve İnceoğlu tarafından İstanbul'da yapılmış bildiri de; Elektronik ticaretteki uygulamalar için vurgulanan veri madenciliği konusunun, bu sektöre kazandıracağı düşünülen faydaları da belirtilmiş, bunun için gerekli teknikler ve yöntemler incelenmiştir[18].

Shen vd., kullanıcıların davranışlarını tahminlemek için e-ticaret sitelerinde gerekli bir yaklaşım sunmuştur[19].

Kalikov A. tarafından, bir yayınevi şirketinin resmi internet sitesi üzerindeki veriler dikkate alınarak yapılmış incelemede, birliktelik kuralları tekniği kullanılarak sipariş ve sepet tablolarındaki kayıtlar incelenmiştir. Kategorilerinin değiştirilmesi gereken ürünler, müşterilerinin ilgi alanları, meslek dağılımları, ilgi alanlarına göre grafikler ve ödeme seçeneklerini de gösteren veri madenciliği uygulaması hayata geçirilmiştir[20].

Yuantao (2008), yaptığı analiz sonucunda, elektronik ticaret verileri üzerindeki hedef müşteri davranışlarını ortaya çıkartmıştır [21].

Feng vd. (2012), elektronik ticarette, ID3 algoritması üzerinde değişiklik yaparak, veri setindeki nitelik seçiminde kullanılan standartları değiştiren ve temelinde müşteri grubu olan bir uygulama hayata geçirmiştir[22].

Zhou & Huang, elektronik ticaret sistemlerinde sınıflandırma tabanlı bir veri madenciliği tekniği üzerine bir araştırma yapmıştır. Bu araştırmada *k-means* algoritması kullanarak,

küme verimliliğini artırmışlardır. Sonucunda ise kullanıcılara yapılan önerilerin hızını artıran bir algoritma ortaya çıkarmış ve sunmuşlardır[23].

Xiong & Cheng (2009), elektronik ticarette veri madenciliği üzerinde durmuştur. Veri madenciliği ile bilgisayar ağ verilerini analiz etmiş ve elektronik ticaretin geleceği hakkındaki gerekli yazılımların geliştirilmesine yönelik frekans bilgilerini sunmuştur[24].

2.2 VERİ MADENCİLİĞİ

Jacobs'a göre, veri madenciliği, ham halde bulunan verinin tek başına sunamayacağı bilgiyi ortaya çıkaran bir veri analizi sürecidir[25].

Türkoğlu ve Doğan, veri madenciliğini, veri yığınları arasından, yakın veya uzak gelecek ile ilgili tahminleme yapabilmemizi sağlayacak bağlantıları bilgisayar uygulamaları yoluyla aranması işi olarak tanımlamıştır[26].

Hand'e göre ise veri madenciliği, veritabanı teknolojisi, istatistik ve makine öğrenme ile yeni bir disiplin ve büyük veritabanları üzerinde önceden tahminlenememiş ilişkilerin ikincil bir analizidir[27].

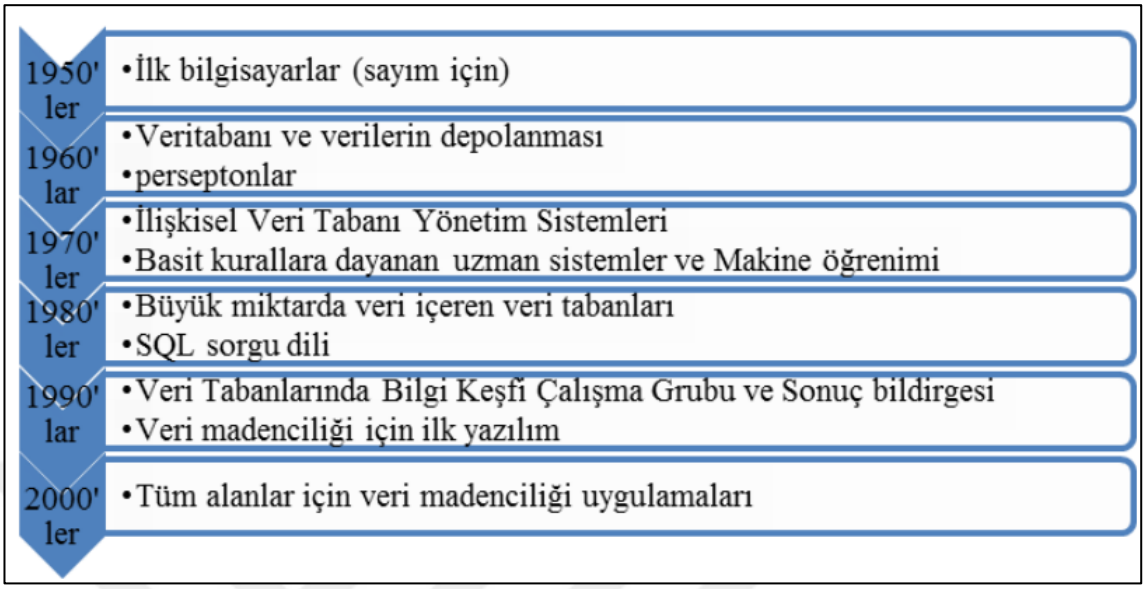
Wang & Kitler için ise veri madenciliği, fazlasıyla tatmin eden ve anahtar niteliğindeki değişkenlerin, yüzlerce potansiyel değişkenden ayrılmasını sağlama yeteneğidir[28].

Tüm bunlar dışında bilgi madenciliği, bilgi keşfi, model/veri analizi, bilgi çıkarımı da veri madenciliği için kullanılmakta olan alternatif terimlerdendir.

2.3 VERİ MADENCİLİĞİNİN TARİHSEL GELİŞİMİ

Günümüzde internet hemen hemen her yerden erişilebilir durumdadır ve yine birçok evde bilgisayar vardır. Disk fiyatlarının düşmesi ve kapasitelerinin artması, bilgiye her yerden ulaşma olasılığı ve kişisel bilgisayarlarda dâhil büyük miktarlarda veriyi saklamayı kolaylaştırmıştır. Geçmişten bugüne kadar veriler yorumlanmış ve bilgi elde etmek istenmiştir. Bu işlemler için ise firmalar üst düzey sistemler oluşturmaya başlamışlardır. Böylelikle bilgi geçmişten bugüne kadar taşınmış oldu. 1950'li yıllarda üretilen ilk bilgisayarlar basit bir hesap makinesi gibi sadece sayım yapmak üzere kullanılmaya başlanmıştır. 60'lı yıllarda ise verilerin depolanması amacı ile veritabanı kavramı, teknoloji dünyasına giriş yapmıştır. 60'lı yılların sonuna gelindiğinde bilim insanları tarafından basit öğrenebilen bilgisayarlar geliştirmişlerdir. Minsky ve Papert, günümüzde sinir ağları adı ile bildiğimiz *Perceptron*'ların yalnızca daha basit ve kolay anlaşılabilen kuralları öğrenebilecek kapasitede olduğunu göstermişlerdir[29] . 1970'li yıllara gelindiğinde İlişkisel Veritabanı sistemleri birçok firma tarafından kullanılmaya başlandı. Bilgisayar konusunda uzman olan insanlar bununla birlikte basit kurallar mantığına dayanan sistemler geliştirmiş ve en temel anlamdaki makine öğrenimine başarmışlardır. 1980'li yıllarda veritabanı yönetim sistemleri daha da yaygınlaşmış ve bilimsel çalışmalarda, mühendislik alanlarında vb. uygulanmaya başlanmıştır. Bu yıllarda firmalar, kendi müşterileri, rekabet içinde olduğu firmalar hakkındaki bilgileri, ürünleri ve hizmetleri ile ilgili verileri içeren veri ambarları oluşturmuşlardır. İçerisinde çok büyük miktarda veri bulunan bu sistemlerdeki verilere SQL veritabanı sorgulama dilini ya da benzer dilleri kullanarak ulaşılabilir. 90'lı yıllara gelindiğinde firmalar ellerindeki çok büyük miktarlardaki veriler içerisinden kendileri için faydalı olan bilgilere erişmek için yollar aramaya başlamışlardır. Bu arayış ile birlikte çalışmalar ve yayınlar oluşmaya başlamıştır. 1992 yılında nihayet veri madenciliği için ilk bilgisayar yazılımı geliştirilmiştir. 2000'li yıllara gelindiğinde, hemen hemen tüm alanlarda uygulanmaya başlayan veri madenciliği aynı zamanda sürekli olarak gelişmiştir. Elde edilen sonuçların faydalı oldukları görüldükçe, bu alandaki ilgi artmıştır. Veri madenciliği ile ilgili tarihsel süreç Şekil 2.1'te gösterilmiştir.

Şekil 2.1: Veri Madenciliğinin Tarihsel Gelişimi



2.4 VERİ MADENCİLİĞİNİN UYGULAMA ALANLARI

Veri madenciliğinin uygulama alanları oldukça geniştir. Veritabanı sistemleri, veri görselliği, yapay sinir ağları, yapay öğrenme ve istatistik gibi disiplinler ve Şekil 2.2'te bazılarına yer verilmiş olan alanlar, veri madenciliği uygulama alanlarına örnektir[30].

Şekil 2.2: Veri Madenciliği ve Disiplinler

2.5 VERİ MADENCİLİĞİNİN FAYDALARI

Veri madenciliğinin kullanımı birçok alanda fayda sağlamaktadır. Bu faydaları şu şekilde sağlamak mümkündür.

a. Müşteri davranışlarının anlaşılmasına ve mevcut müşterilerin elde tutulmaya devam edilmesine yardımcı olur.

b. Kredi risk davranışlarına uygun modelleri finans sektörü için oluşturarak, yeni yapılan başvurulardaki müşterilerin riskini minimuma indirir.

c. Borsa ve finans sektöründe stok fiyatlarının tahminini ve borsanın yönetimini yapar.

d. Finans sektöründeki firmalar için mevcut müşterilerin ödeme performanslarının incelenip, bu performansları kötü olan müşterilerin tespitini, ortak özelliklerinin belirleyerek yapılması ve en önemlisi bu müşterilere özel risk yönetim politikası geliştirilmesinde kullanılır.

e. En iyi veya en kötü müşteri bölümlerinin bulunmasını sağlayarak bu müşteriler için yeni pazarlama adımları atılmasını kolaylaştırır.

f. Firmaların düzenlemiş olduğu hizmet veya ürün kampanyalarında mevcut müşterilerin kitlesinin belirlenmesi ve bu müşterilerin davranışlarına özgü kampanyaların oluşturulmasını kolaylaştırır.

g. Sağlık alanlarında, hastalara yapılan testler veya tarama testleri sonucunda elde edilen verilerin kullanılarak çeşitli hastalıkları için ön tanı yapılması, kalp krizi riskleri belirlenmesi ve acil servislerdeki hastaları belirtilerine göre önceliklendirilmesi sağlanabilir.

h. Bankacılık, sigortacılık ve telekomünikasyon alanlarında elde var olan veriler incelenerek sahtekârlık girişimlerini tespit etmek veya buna benzer davranış gösterenleri belirlemek amacı ile kullanılabilir.

3. VERİ VE YÖNTEM

3.1 VERİ MADENCİLİĞİNİN BİLGİ KEŞFİ SÜRECİ

3.1.1 Verinin Temizliği

Bu adım tutarsız, gürültülü veya eksik verilerin belirlenerek bu verilerin temizlenmesi sürecidir.

3.1.2 Verini Bütünleştirmesi

Bir veya daha fazla veri kaynağından alınmış verilerin birleştirilmesi sürecidir.

3.1.3 Verinin Seçilmesi

Veri tabanından analiz edilerek alınmış verilerden, problemin çözümü için bize gerekli olan verinin seçilmesi sürecidir.

3.1.4 Verinin Dönüştürmesi

Veri dönüştürme aşamasında verinin madencilikte kullanılabilecek hale gelmesi için verinin uygun formata dönüştürülmesi işlemi yapılır.

3.1.5 Veri Madenciliğinin Uygulaması

Birçok süreçten geçerek hazırlanmış veri üzerinde, veri madenciliğinde kullanılan algoritmaların uygulanması sürecidir.

3.1.6 Veri Desenleri

Bazı kriter ve ölçümlere göre elde edilen veriyi temsil edecek olan örüntülerin tanımlanması sürecidir.

3.1.7 Bilginin Sunumu

Veri madenciliği süreci uygulanarak elde edilen bilgilerin kullanıcıya gösterildiği aşamadır.

3.2 VERİ MADENCİLİĞİ YÖNTEMLERİ

3.2.1 Veri Madenciliğindeki Birliktelik Kuralları

Bu yöntem veri madenciliğinin en iyi bilinen yöntemlerinden biridir. Çok büyük veri ambarları içerisindeki büyük miktardaki veriler arasından birbiri ile ilişkisi olan değişkenlerin tespiti ve aralarındaki bağlantının boyutunu bulmak kullanılır. *Sequence, Eclat, GRI, Carma* ve *Apriori* bu yöntemlerde kullanılan bazı algoritmalardandır.

Örneğin; Süpermarket veritabanı incelenerek müşteriler alışveriş alışkanlıklarına göre analiz edildiğinde, çocuk bezi alan müşterilerin yüzde kaçı aynı zamanda salça alıyor veya her alışverişinde mutlaka süt ve makarna alan müşterilerin yüzde kaçı sebze reyonundan alışveriş yapıyor gibi sonuçları ortaya koymak kolaydır.

Örneğin; Alışveriş alışkanlıkları incelenerek çocuk bezi alanlar yüzde kaç oranında salça veya makarna ve süt alanların yüzde kaçı sebze alıyor gibi ilişkilerin tespit edilmesi mümkündür.

Çocuk Bezi → | yüzde 43 (satılan toplam ürünlerin yüzde 43'ünde çocuk bezi alınıyor.)
Çocuk Bezi → Salça | yüzde 45 (Çocuk Bezi alanlar yüzde 45 oranında süt de alıyor.)
Süt, Makarna → Sebze | yüzde 40 (Süt ve makarna alanların yüzde 40 oranında sebze alıyor)

3.2.2 Tahmin ve Sınıflandırma

Gelecekte oluşacak verilerin hangi yönde eğimde olacağını tahmin edebilmek için nesnenin özelliklerinin incelenmesi ve önceden tanımlanmış olan bir sınıfa atama işlemidir. *Random Forest, KNN, Decision Tree, Navie Bayes* sınıflandırma yöntemlerinde en çok kullanılan algoritmalardan bazılarıdır.

Örneğin; Geçmiş hastalık verilerinin incelenerek yeni oluşabilecek hastalığı teşhisi, kullanıcının davranışlarını belirleme, Ses tanıma, kredi başvurusunda bulunacak müşterinin riskinin analiz edilerek kredi verilebilirliği tespiti birer sınıflandırma örneği olarak gösterilebilir.

3.2.3 Kümeleme Analizi (Kümeleme Yöntemi)

Bu yöntemin kullanılmasının asıl amacı, veri tabanında dağınık şekilde duran verilerin, ayırt edici özelliklerine göre birleştirilerek daha sonradan işlenebilir hale getirilmesidir. Sınıflandırma yöntemine benziyor olsa da aradaki fark birleştirilecek kümelerin önceden belirlenmiş olması. Bu yöntem müşterilerinizin profillerini oluşturmak için de kullanılabilir. *K-metoids* ve *k-means* bu amaçla kullanılabilecek bazı algoritmalardır.

Örneğin; Biyoloji alanındaki bitki ve hayvanların sınıflandırılması, şehir planlaması yapılırken evlerin değerleri, konumları ve hatta tiplerine göre ayrılması, marketlerdeki farklı müşteri gruplarının belirlenmesi ve alışveriş özelliklerine göre ortaya konması kümeleme örneklerindendir.

3.2.4 Aykırılık Analizi

Verilerin analiz yöntemleri ile kontrol edilip aykırı değerlerin veya aşırı sapma olan değerlerin bulunması sürecidir. Bu aykırı veriler, verilerin ölçülmesi, kaydedilmesi veya okunması sırasında gerçekleşen hatalardan kaynaklanmaktadır. Bu yöntem, işte bu aykırı verileri en aza indirerek çıkacak sonucun daha sağlıklı olmasını amaçlamaktadır.

Hawkins (1980), Aykırı veriyi şöyle açıklıyor; "Bir aykırı değer, farklı bir mekanizma tarafından yaratıldığı şüphelerini uyandırmak için diğer gözlemlerden çok sapan bir gözlemdir"[31].

Örneğin; Banka hesaplarındaki kısa süre içerisindeki çok fazla işlemin yapılmasının tespitinin yapılarak dolandırıcılık vakalarının tespiti sağlanabilir, ya da sağlık sektöründeki hastalara uygulanan tedavi ve ilaç sayılarının tespiti yapılabilir. Hastalar ve kullandıkları ilaçlar üzerine bir analiz yapılacağı zaman, baş ağrısı olan biri ile kanser hastası olan birisinin kullandıkları ilaçların sayısı aynı olmayacaktır. Kanser hastalarının ilaç sayıları çok yüksek olacağı için bu veri analiz dışında bırakılabilir.

3.3 VERİ TEMİZLEME

Veri temizleme ile ele alınan veri kalitesi sorunlarını sınıflandırır ve ana çözüm yaklaşımlarına genel bir bakış sunarız. Veri temizleme özellikle heterojen verileri entegre ederken gereklidir. Şema ile ilgili veri dönüşümleri ile birlikte ele alınmalıdır. Veri ambarlarında, veri temizleme *ETL* sürecinin önemli bir parçasıdır.

Veri temizleme, verilerin kalitesini artırmak için verilerdeki hata ve tutarsızlıkların tespiti ve kaldırılması ile ilgilenir. Veri kalitesi sorunları, örneğin veri girişi sırasında yazım hataları, eksik bilgiler veya diğer geçersiz veriler nedeniyle, dosyalar ve veritabanları gibi tek veri koleksiyonlarında bulunur. Birden fazla veri kaynağının, örneğin veri

ambarlarına, birleşik veritabanı sistemlerine veya küresel web tabanlı bilgi sistemlerine entegre edilmesi gerektiğinde, veri temizleme ihtiyacı önemli ölçüde artar. Bunun nedeni, kaynakların genellikle farklı durumlarda gereksiz veriler içermesidir. Doğru ve tutarlı verilere erişim sağlamak için farklı verilerin birleştirilmesi mükerrer bilgilerin tespiti ve ortadan kaldırılması gerekli hale gelir.

Veri ambarları, veri temizliği için kapsamlı destek gerektirir ve sağlar. Çeşitli kaynaklardan çok miktarda veri yükler ve sürekli olarak yeniler, böylece bazı kaynakların "kirli veri" içermesi olasılığı yüksektir. Ayrıca, veri ambarları karar vermede kullanılır, böylece yanlış sonuçlardan kaçınmak için verilerinin doğruluğu hayati önem taşır. Örneğin, yinelenen veya eksik bilgiler yanlış veya yanıltıcı istatistikler üretir. Çok çeşitli olası veri tutarsızlıkları ve veri hacmi nedeniyle, veri temizliği, veri depolamadaki en büyük sorunlardan biri olarak kabul edilir.
Veri temizliği tipik olarak, dönüştürülmüş verileri depoya yüklemeden önce ayrı bir veri hazırlama alanında gerçekleştirilir. Bu görevleri desteklemek için çok sayıda işlevselliğe sahip çok sayıda araç mevcuttur, ancak temizleme ve dönüştürme işlerinin önemli bir bölümünün elle yazılması veya bakımı zor olmayan düşük seviyeli programlar tarafından yapılması gerekir.

3.3.1 Veri Temizleme Yaklaşımları

Genel olarak, veri temizleme birkaç aşamadan oluşur[32] ;

a. **Veri analizi**: Hangi tür hataların ve tutarsızlıkların kaldırılacağını tespit etmek için ayrıntılı bir veri analizi gereklidir. Veri veya veri örneklerinin manuel olarak incelenmesine ek olarak, veri özellikleri hakkında meta veri elde etmek ve veri kalitesi sorunlarını tespit etmek için analiz programları kullanılmalıdır.

b. **Dönüşüm iş akışının ve eşleştirme kurallarının tanımlanması:** Veri kaynaklarının sayısına, bunların heterojenlik derecesine ve verilerin "kirliliğine" bağlı olarak, çok sayıda veri dönüşümü ve temizleme adımının gerçekleştirilmesi gerekebilir. Erken veri temizleme adımları, tek kaynaklı örnek sorunlarını düzeltebilir ve verileri entegrasyon için hazırlayabilir. Sonraki adımlar şema verilerle ilgilenir. Veri depolama için, bu dönüştürme ve temizleme adımlarının kontrol ve veri akışı, *ETL* sürecini tanımlayan bir iş akışı içinde belirtilmelidir. Şema ile ilgili veri dönüşümleri ve temizleme adımları, dönüşüm kodunun otomatik olarak oluşturulmasını sağlamak için mümkün olduğunca anlaşılabilir bir sorgu dili ile belirtilmelidir. Ayrıca, veri dönüştürme iş akışı sırasında kullanıcı tarafından yazılmış temizleme kodunu ve özel amaçlı araçları kullanmak mümkün olmalıdır.

c. **Veri Doğrulama:** Gerekirse tanımları iyileştirmek için bir dönüşüm iş akışının ve dönüşüm tanımlarının doğruluğu ve etkinliği, kaynak verilerin bir örneği veya kopyası üzerinde test edilmeli ve değerlendirilmelidir. Analiz, tasarım ve doğrulama adımlarının çoklu iterasyonlarına ihtiyaç duyulabilir, çünkü bazı hatalar ancak bazı dönüşümler uygulandıktan sonra belirginleşir.

d. **Veri Dönüştürme:** Bir veri ambarını yüklemek ve yenilemek için *ETL* iş akışını uygulayarak veya birden çok kaynaktaki sorguları yanıtlarken dönüştürme adımlarının yürütülmesi aşaması.

e. **Temizlenen verilerin geri akışı:** Hatalar giderildikten sonra, eski uygulamalara eski verileri de iyileştirmek ve gelecekteki veri çıkarma işlemleri için temizleme işinin yeniden yapılmasını önlemek için temizlenen veriler orijinal kaynaklardaki kirli verilerin de yerini almalıdır. Veri depolama için, temizlenen veriler veri hazırlama alanından elde edilebilir[33].

3.4 VERİ DÖNÜŞTÜRME

Kullanılan istatistiki model ham veri üzerinde çeşitli dönüşümler yapılmasını gerekli kılabilir. Bunun sebeplerinden biri de değişkenlerin ölçüldüğü birimlerin çok büyük farklılık gösterebilecek olmasıdır. Örneğin bir cep telefonu abonesinin aylık faturasının TL cinsinden değeri ve attığı SMS sayısı çok farklı ortalama ve varyansa sahip olabilir. Bu ölçek farkları kullanılan veri madenciliği algoritmasında bir değişkenin sonuca olması gerektiğinden fazla etki etmesini sağlayabilir. Değişkenleri "tek tip" hale getirmek için birçok standardizasyon yöntemi mevcuttur. En sık kullanılan standartlaştırma yöntemleri *min-max* ve z-skor standartlaştırma yöntemleridir. *Min-max* standartlaştırma yöntemi sadece minimum ve maksimum değere göre minimum değer 0 maksimum değer de 1 olacak şekilde veriyi standardize eder.

Yani,

$$X^* = \frac{X - \min(X)}{aralik(X)} = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{3.1}$$

Burada *min-max* standart değerlerin verideki uç değerlere karşı hassas olduğunu belirtmek gerekir. Ayrıca veri aşağı yukarı tekdüze bir dağılıma sahipse *min-max* standart değerleri yaklaşık olarak yüzdeliklere denk gelecektir. Bir örnek vermek gerekirse 2005 *Capital* 500 listesindeki 500 şirketin cirolarına bakıldığında en yüksek 10.364.153.845 TL ile Petrol Ofisi, en düşük ise 63.082.552 TL ile Rast Gıda olarak görülmektedir. *Min-max* standardizasyonu sonucu en düşük değer 0 en yüksek ise 1 değerini alacaktır. Hemen birinci şirketten sonra gelen şirket olan Ford Otosan ise 0,53 değeri almaktadır ki en yüksek değerin yarısı kadar ciro yaptığını gösterir. Aynı işlem çalışan sayısı değişkenine göre yapıldığında en fazla çalışan sayısına sahip şirketin 7539 adet çalışana, en düşük çalışan sayısına sahip şirketin de 14 çalışana sahip olduğu görülmektedir. Bu durumda örneğin 710 çalışana sahip Pınar Süt 0,09 değerine sahip olacaktır.

Z-skor standartlaştırması ise değişkeni ortalama değeri 0 ve standart sapma değeri 1 olan bir değişkene dönüştürür. Bunu yapmak için ortalamayı çıkarıp standart sapmaya bölmek yeterlidir. Burada "-" değerler ortalamadan alttaki gözlemleri "+" değerler ise ortalamadan yüksek gözlemleri gösterir. Ortalamadan fark ise standart sapma biriminden ölçülür. Z değeri şu şekilde hesaplanır:

$$X^* = \frac{X - ortalama(X)}{s\tan dartsapma(X)} = \frac{X - \overline{X}_x}{s_x} \tag{3.2}$$

Bir örnek vermek gerekirse 35 kişilik bir sınıfta sınav ortalamasının 63,09 standart sapmasının ise 13,77 olduğunu düşünelim. Bu durumda en yüksek not olan 94 z değeri olarak 2,24'e denk gelecektir. En düşük değer olan 35 ise -2,04 z değerine denk gelecektir[34].

3.5 VERİ İNDİRGEME

Analiz için bütün veri ambarından veri seçtiğinizi düşünün. Veri kümesi büyük olasılıkla büyük olacaktır! Bu çok büyük miktardaki veri üzerinde karmaşık veri analizi ve veri madenciliği uzun zaman alabilir, bu da bu analizi pratik veya olanaksız hale getirir.

Veri indirgeme veya küçültme teknikleri, hacim olarak çok daha küçük olan ancak sahip olunan asıl verilerin bütünlüğünü bozmadan veri setinin daha düşük bir temsilini elde etmek için uygulanır. Yani, indirgenmiş veri setindeki veri madenciliği süreci daha verimli olmalı, fakat aynı (veya çok yakın) analitik sonuçları üretmelidir.

3.6 KÜMELEME ANALİZİ

İnsanlar hayatları ile ilgili olsun veya olmasın önemli bir karar almadan önce bir geri adım atarak büyük resmi görmek isterler. Ancak bazı zamanlarda bu görmeyi istediğimiz büyük resim anlaşılmayacak derecede karmaşık olabilir. Çok büyük ve karmaşık sorunların çözümünde izlenen yol genellikle, büyük sorunları daha küçük parçalara ayırmak veya tek başına rahat çözülebilecek alt sorunlara ayırmaktır. Daha sonra çözüme kavuşturulmuş alt sorunlar birleştirir ve sonuca gidilir. Ancak bazı senaryolarda veriler öyle dağınık halde olurlar ki bölme ve alt gruplara ayırma işlemine nereden başlayacağınızı bilemezsiniz. İşte tam da bu tür durumlar için kümeleme algoritmaları geliştirilmiştir. Taksonomi veya Segment analizi olarak da bilinen kümeleme analizlerinin genel amacı, gruplandırılmamış verileri benzer özelliklerine göre sınıflandırarak araştırmacıya işe yarar ve özetleyici bilgi elde etmesine yardımcı olmaktır diyebiliriz.

Kümeleme analizi veri gruplarını nesne olarak kabul eder. Nesneleri gruplara veya kümelere ayırırlar, böylece bir kümedeki nesneler birbirine "benzer" ve diğer kümelerdeki nesnelere "benzemez". Benzerlik, bir uzaklık fonksiyonuna bağlı olarak, nesnelerin uzayda ne kadar "yakın" oldukları ile tanımlanır. Bir veri kümesinin "kalitesi" çapı, kümedeki iki obje arasındaki en uzak mesafe ile temsil edilir. Sentroid mesafesi, veri seti kalitesinin alternatif sayılabilecek bir ölçüsüdür ve her veri seti nesnesinin küme sentroid'inden ortalama mesafesi olarak tanımlanır.

3.6.1 K-means Kümeleme Algoritması

K-means ismiyle daha çok bilinen "K-Ortalamalar" yöntemi bir "*Unsupervised Learning*" ve kümeleme algoritmasıdır. Bu algoritmadaki deki **K değeri kümelenmesi gereken sayıyı** temsil eder. Bu değer algoritmanın çalışabilmesi için parametre değişkeni olarak verilmesi gerekir. Bu durum aslında bir dezavantajdır. K değerini kendi hesaplayan ***x-means*** adında başka bir algoritmada vardır. Algoritmanın basit bir çalışma şekli vardır. K değeri belirlendikten sonra algoritmada rastgele "**K**" tane merkez noktası belirleyerek

işe başlanır. Set içerisindeki her veri ile en başta rastgele belirlenmiş olan merkez noktaları arasındaki uzaklık hesaplanır ve veri en yakındaki merkez noktasına bakılarak o kümeye atar. Daha sonra bu kümeler için yeniden bir merkez noktası belirlenir. Ardından yeni merkez noktaları baz alınarak yeniden kümeleme işlemi yapılır. Bu süreç sistem en kararlı hale gelene işletilir. *K-means* algoritmasındaki başlangıç merkez noktalarının rastgele atanması sorunlara yol açabilmektedir. Başlangıç noktalarını daha iyi seçebilmek için 2007 yılında **David Arthur** ve **Sergei Vassilvitskii** *k-means* algoritmasının bir varyasyonu olan *k-means*++ algoritmasını geliştirilmiştir. ***k-means*++**'a göre rastgele başlangıç noktası seçildikten sonra diğer tüm veriler ile arasındaki mesafe hesaplanır. Bu mesafenin karesi alınıp belli hesaplamalar yaparak yeni başlangıç noktaları seçilir. Bu işlemin karmaşıklık analizi **O(log k)**'dır.

3.6.2 Birch Kümeleme Algoritması

BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*), T. Zhang, Ramakrishnan ve Livny tarafından, büyük miktardaki veriler için geliştirilmiş bir kümeleme algoritmasıdır. Bu algoritma alanında önerilen ilk algoritmadır. *Birch* algoritması en temelde hiyerarşik kümeleme analizi yapılırken karşılaşılan önceki adıma dönmenin mümkün olmadığı ve ölçekleme sorunlarına çözümler getirmiştir. Bahsi geçen büyük hacimli veri setlerinde kümeleme işlemi uygulanırken *NxN* boyutlu bir uzaklık matrisi kullanıldığından, alan karmaşıklığı problemi ortaya çıkmaktadır. *Birch* algoritması ile *NxN* boyutunda uzaklık matrisi yerine kümelerin niteliklerini temsil eden bir ağaç oluşturulur ve bu oluşturulan ağaç üzerinden kümeleme yapılarak alan karmaşıklığı problemi ile karşılaşılmaz[35].

Nitekim Tong ve Kang (2014, s.261), BIRCH algoritmasını, sınırlı hafıza ile baş etmenin bir yolu olarak aktarmıştır[36].

3.7 VERİ ANALİZ İŞLEMLERİ

Analizde kullanılan veri, bir turizm firmasının son altı yıllık müşteri ve rezervasyon bilgilerinin bir araya getirilmesiyle ilişkisel veri tabanı tablosu yapısında oluşturulmuştur. Rezervasyon bilgisi müşteri bazında gruplanarak müşteri bilgisine eklenmiştir. (Veri, kişisel verilerin korunması kanunu (KVKK) gereğince isim, soy isim, cep telefonu, e-posta gibi kişiyle doğrudan ilişkilendirilebilecek özniteliklerden arındırılmıştır.) Veri kaynaktan alınırken bir *ETL* ve veri işleme yazılımı olan ve büyük veri kümesi üzerinde çalışan Datameer kullanılmıştır. Turizmde bulunan rezervasyon verisi telekomünikasyon ya da banka verileriyle karşılaştırıldığında daha kirli ve eksik halde bulunmaktadır. Bunun sebebi, müşterilerinin genel olarak sadece doldurmak zorunda oldukları alanlarla ilgili bilgi vermesi ve bu bilgilerin doğruluğunun bahsedilen diğer sektörlerdeki kadar kritik olmamasıdır. Veri önişleme adımının yapılması analizler üzerinde olumlu yönde, önemli ölçüde etki etmektedir. Önişleme sonucunda verinin temizlenmesiyle, yapılacak analiz daha tutarlı ve anlamlı hale gelmektedir. Bu yüzden veri analizde kullanılmadan önce bazı önişleme adımlarından geçirilmiştir. Özniteliklerdeki boş değerler hesaplanmış, yüzde 70'nin üzerinde boş olan kolonlar atılmış. Bu ortalamanın altındaki veriler kendi kolonundaki diğer değerlerin ortalaması ile doldurulmuştur. Kategorik veri içeren ürün tiplerinin bulunduğu kolon üzerinde ikili(*binary*) sistem (*one-hot encoding*) çalışması yapılarak değerler farklı kolonlarda {0, 1} değerlerini alacak şekilde ifade edilmiştir.

Aykırı değer, uzayda gözlemlenen bir noktanın, gözlemlenen diğer noktalardan uzak olması durumunda ortaya çıkar ve özellikleriyle kümenin diğer elemanlarından ayrılır. Aykırı değerler oluşmasının en önemli sebebi, veri çeşitliliği ile beraber, veriler kaydedilirken ya da toplanırken hata yapılmasıdır. Aykırı değerlerin fazla olması veri setinin normal dağılımdan sapmasına ve yapılacak analizlerin olumsuz yönde etkilenmesine sebep olabilir. Bu adımda önemli olan hata sonucu oluşmuş değerlerin ihmal edilmesidir. Diğer değerlerden farklı olmasından dolayı aykırı olarak belirlenen değerlerin, gerçekten aykırı olup olmadığı kontrolü sağlanmalıdır. Z-değeri sonucunda veri setinin dağılımı normal dağılıma daha çok benzetilerek sonucun daha güvenilir ve tutarlı olması sağlanmıştır.

Çalışmada aykırı değerlerin tespitinde z-değeri yöntemi kullanılmıştır. Z-değeri yönteminde, verilerin, veri gruplarıyla standart sapma ve ortalama açısında ilişkisi incelenir. Z-değeri sonucunda değerler aynı ölçek, yani 0 ortalama, 1 standart sapma üzerine gelir.

3.7.1 Verinin Kaynaktan Yüklenmesi

Bu süreçte farklı veri kaynaklarından toplanarak. xlsx uzantılı dosya üzerine aktarılmış olan 200000 satır veri analiz edilmek üzere uygulama üzerine yüklenecek. Bu kayıtlar analiz işlerini yapacağımız *Python* programlama diline, *Pandas* kütüphanesi yardımı ile aktarılarak veri seti olarak saklanacak. Verinin görselleştirilmesi, *k-means*, Z-değeri gibi algoritmaların da kullanılabilmesi için kullandığımız kütüphaneler arasına, "*numpy*", "*matplotlib*", "*seaborn*", "*sklearn*" ve "*pandas*" eklenmiştir. İlgili kütüphanelerin eklenmesi ve verinin veri seti olarak içeri alınmasına ait kod örneği Şekil 3.1'te verilmiştir.

Şekil 3.1: Python programlama diline kütüphanelerin eklenmesi

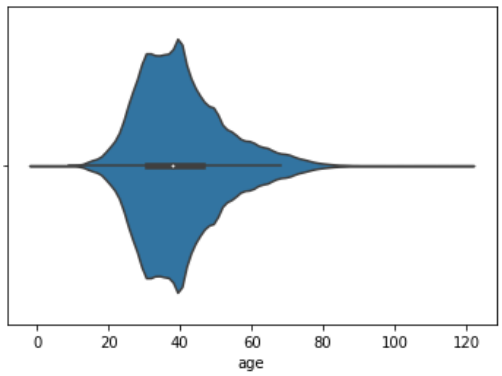
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.cluster import KMeans
from sklearn import preprocessing

df = pd.read_excel('data.xlsx')
```

3.7.2 Aykırı Verinin Temizlenmesi

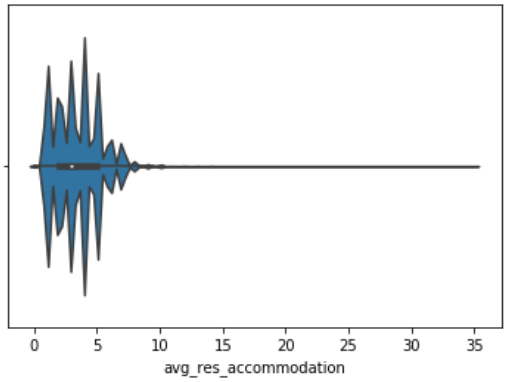
Veri seti haline getirilmiş 200000 satır veri üzerindeki aykırı verilerin tespit edilerek analiz yapılırken sonuçların doğruluk paylarının azalmasına sebep olacak kayıtlar bu

süreçte temizleniyor. Aykırı verilerin belirlenmesi işlemi sadece diğer bir sütundaki veri ile ilişkisi olmayan sütunlar için belirlenmelidir. Örneğin; rezervasyonlara ödenen toplam para değişkeni, rezervasyon sayısı ile doğru orantılıdır. Rezervasyon sayısı arttıkça ödenen ücretin de artması beklenir. Aynı şekilde rezervasyon başına geçirilen ortalama geceleme sayısı da tek başına bir değerdir ancak yaş değeri diğer alanlardan bağımsız olarak değişebilir. 50 rezervasyonu olan misafir ile tatillerinde en çok yurtdışını tercih eden misafirlerin yaşları aynı veya benzer olabilir. Dolayısı ile yaş veya ortalama geceleme rakamları aykırı veri olarak temizlik aşamasında işlenebilir. Ancak, toplam ödenen tutar alanında aykırı veri tespiti yapmak mantıklı olmayacaktır çünkü rezervasyon sayısı ile doğru orantılı olan bu alanın rezervasyon sayısı değişkeni ile bir bağlantısı vardır. Şekil 3.2 ve 3.3 de aykırı veriler temizlenmeden önceki "Yaş" ve "Rezervasyon başına ortalama geceleme sayısı" bilgisine ait grafik yer almaktadır. Aykırı veriler temizlendikten sonraki hallerini ise Şekil 3.4 ve 3.5'da görebilirsiniz. Aykırı veriler temizlendikten sonra en başta 200000 olan satır sayımız 175144 olmuştur. Toplamda 24856 adet aykırı veri temizlenmiş oldu.

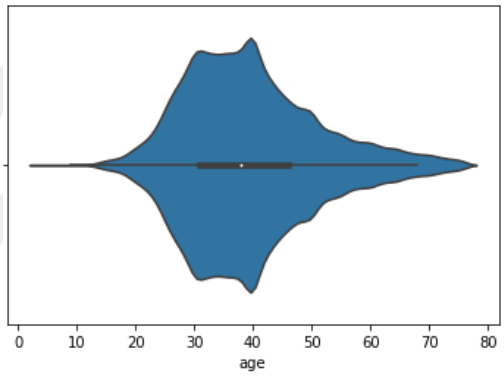
Şekil 3.2: Temizlik öncesi yaş verisi



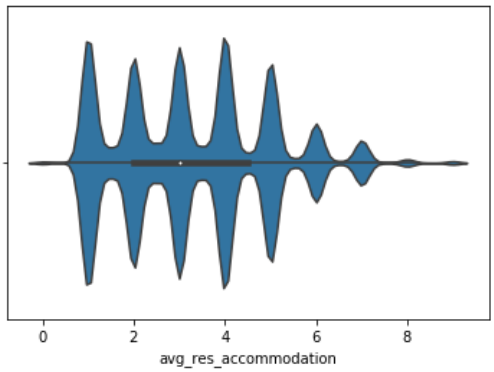
Şekil 3.3: Temizlik öncesi konaklama sayısı



Şekil 3.4: Temizlik sonrası yaş verisi



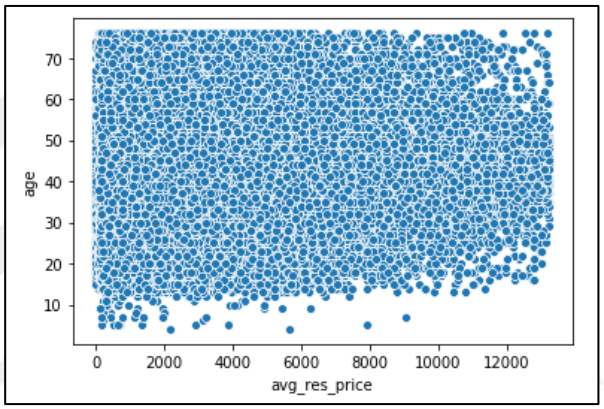
Şekil 3.5: Temizlik sonrası konaklama sayısı



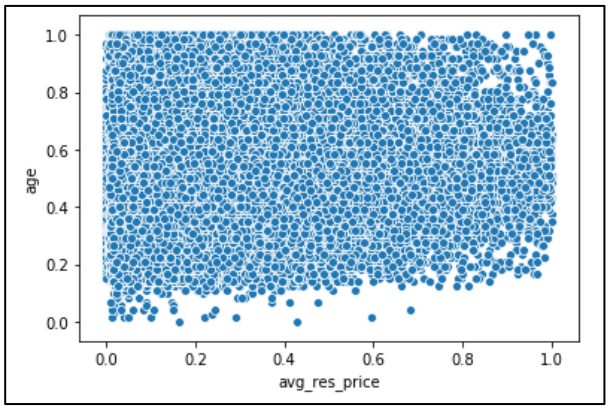
3.7.3 Verinin Dönüştürülmesi

Elimizde aykırı veriler temizlendikten sonra kalan 175144 satır veri üzerinde *min-max* normalizasyonu uygulayarak satırlar içerisindeki tüm verileri 0-1 aralığına çekeceğiz. Normalizasyon işlemi yapılmadan önce "Yaş" ve "Rezervasyon Başına Ödenen Ücret" grafiği Şekil 3.6'daki gibiyken, normalizasyon sonrası Şekil 3.7'deki halini almıştır.

Şekil 3.6: Normalizasyon öncesi



Şekil 3.7: Normalizasyon Sonrası



3.7.4 Küme Sayısının Elbow Yöntemi Kullanılarak Bulunması

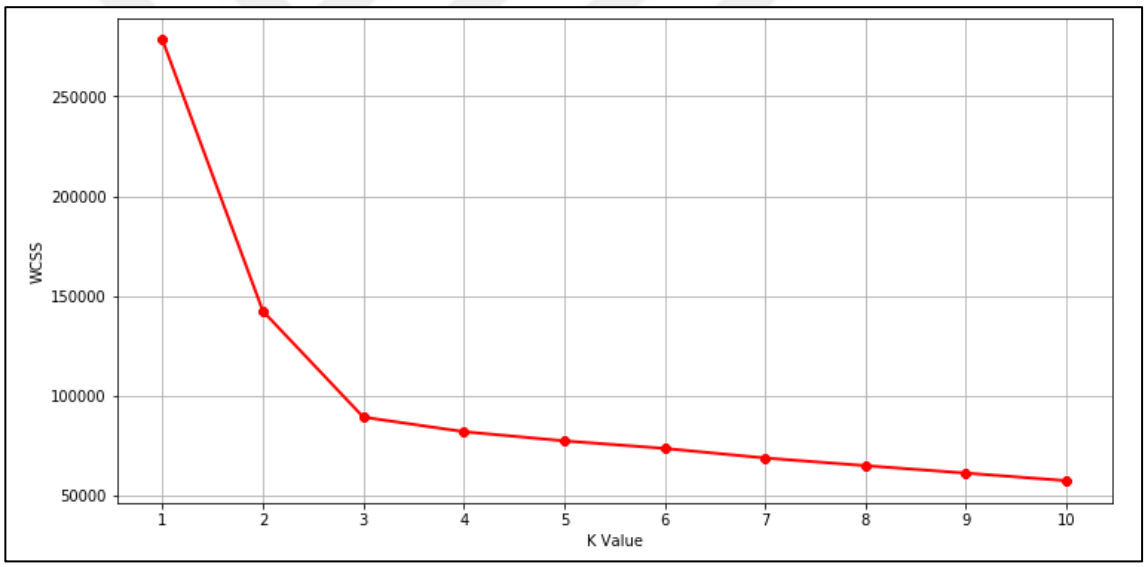
Verilerin kümelenmesi için kullanacağımız k-means, kendisine parametre olarak verilen bir veri setini yine kullanıcı tarafından belirlenmiş olan sayıda (k) kümeleyen bir veri madenciliği algoritması olarak biliniyor. Bu algoritma veri setini analiz ederek k sayısı kadar küme olmasa bile veriyi kümelemeye çalışır. Dolayısı ile algoritmanın doğru sonucu vermesi için kaç farklı kümeye ayırılmasını biliyor olmak gerekiyor. İşte tam bu noktada *Elbow* yönteminin bulunmasının önemi ortaya çıkıyor. *Elbow* yöntemindeki ana fikir, *k-means* algoritmasına sıra ile k değerleri göndererek kümeleme yapmasını sağlamak ve daha sonra her bir k değeri için küme içindeki noktaların algoritmanın belirlemiş olduğu küme merkezlerine olan uzaklıklarının karelerini alarak bunun da ortalamasını hesaplamaktır. (*Within Cluster Sum of Squares*) Kod Örneği Şekil 3.8'de verilmiştir.

Şekil 3.8: Elbow yöntemi kod örneği

```
var sse = {};
for (var k = 1; k <= maxK; ++k) {
 sse[k] = 0;
 clusters = kmeans(dataset, k);
 clusters.forEach(function(cluster) {
 mean = clusterMean(cluster);
 cluster.forEach(function(datapoint) {
 sse[k] += Math.pow(datapoint - mean, 2);
 });
 });
}
```

Daha sonra k'nin her bir değer için *WCSS* grafiği çizilir. Çizilmiş olan grafiğin bir kol gibi göründüğü varsayılırsa dirsek noktası bizim için en iyi k değeridir. Buradaki ana fikir, küçük bir *WCSS* noktası bulmaktır ancak *WCSS* değeri, k değer arttıkça 0'a düşme eğilimi gösterir. Bu sebeple amacımız düşük *WCSS* değerine sahip küçük bir k değeri belirlemek. Dirsek noktası genellikle k'yı artırarak azalan geri dönüşlere başlanmış yeri temsil ediyor. *Elbow* yöntemini veri setimiz üzerinde çalıştırıp grafiğe dönüştürdüğümüzde Şekil 3.9'teki gibi bir grafik çıkıyor ortaya. Grafik gösteriyor ki veri içerisinde kümeleme yapılması için belirlenmiş olan uygun küme sayısı 3'tür.

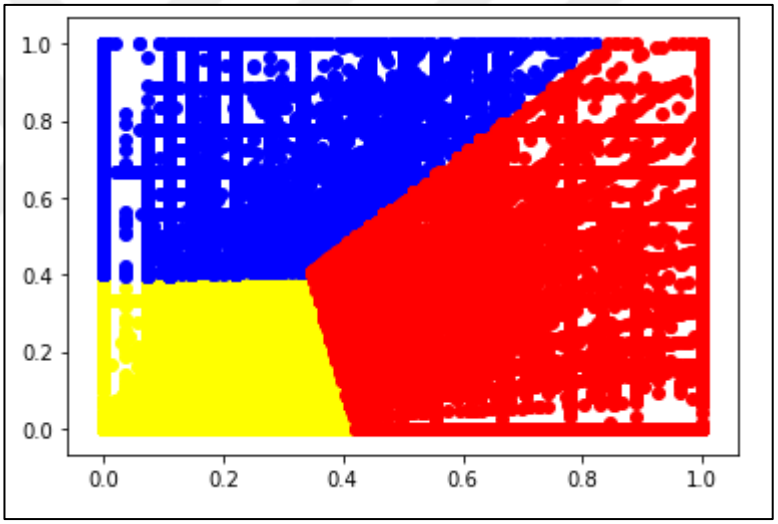
Şekil 3.9: Elbow yöntemi uygulanan veri grafiği



3.7.5 Uygulama ve Sonuçlar

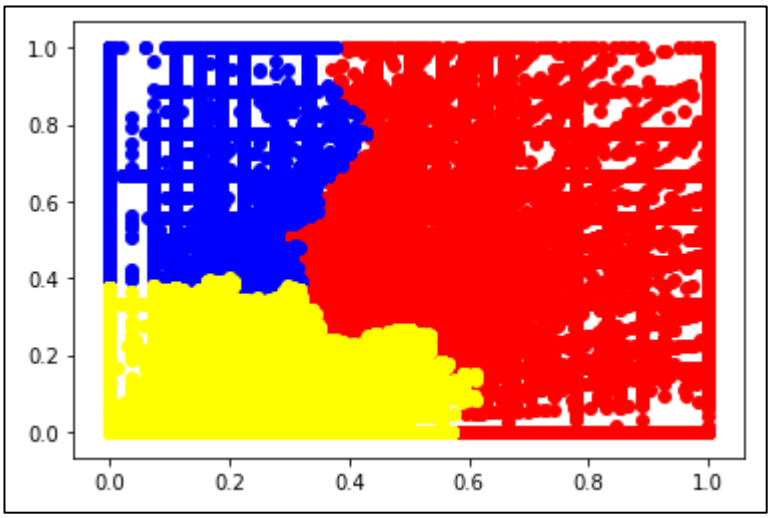
K-means bizim için elimizdeki verinin kümelenmesi işlemini yapacaktır. *Elbow* yöntemini kullanarak optimum küme sayısının 3 olduğunu belirlemiştik.(Şekil 3.9) Şimdi *k-means* algoritmasını bu parametre ile, birbiri ile ilişkileri dışarıda bıraktığımız verimiz için çalıştıralım. Sonuç Şekil 3.10'da görüldüğü gibi karşımıza çıkıyor.

Şekil 3.10: K-means uygulanmış verilere ait kümeler



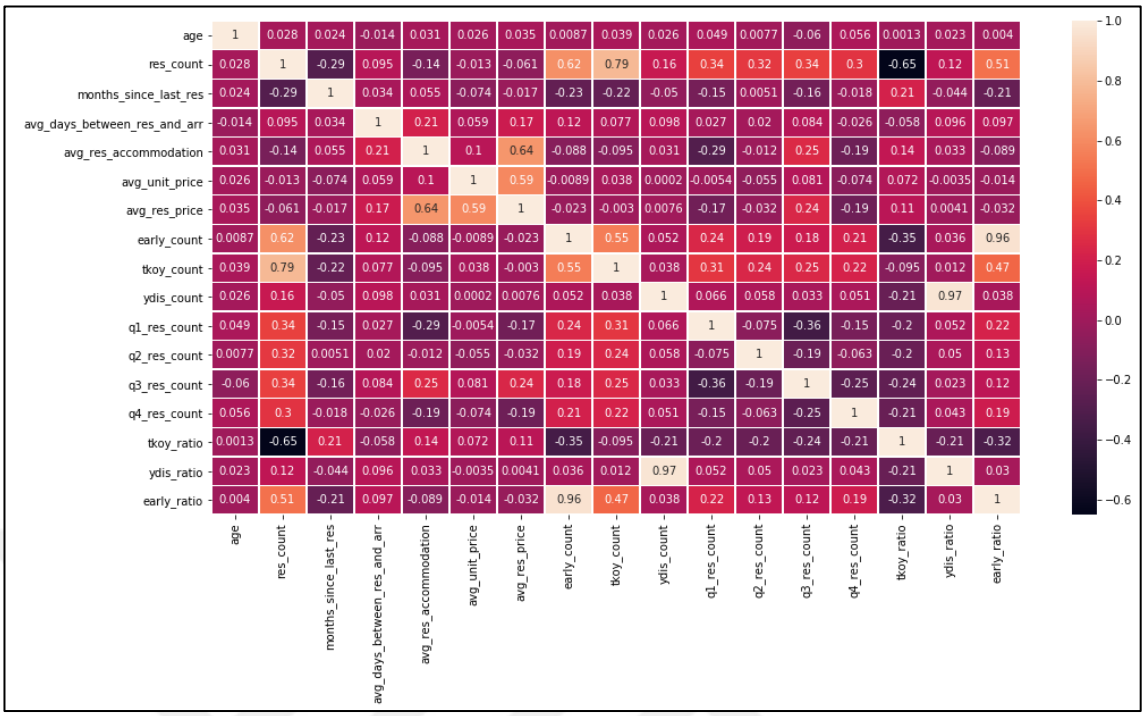
Aynı veri seti üzerine "*k-means"*'den başka bir kümeleme algoritması olan *"Birch"* algoritmasını uygulayarak kümelenmiş verileri daha net şekilde Şekil 3.11'de görülebilir.

Şekil 3.11: Birch yöntemi uygulanmış verilere ait kümeler

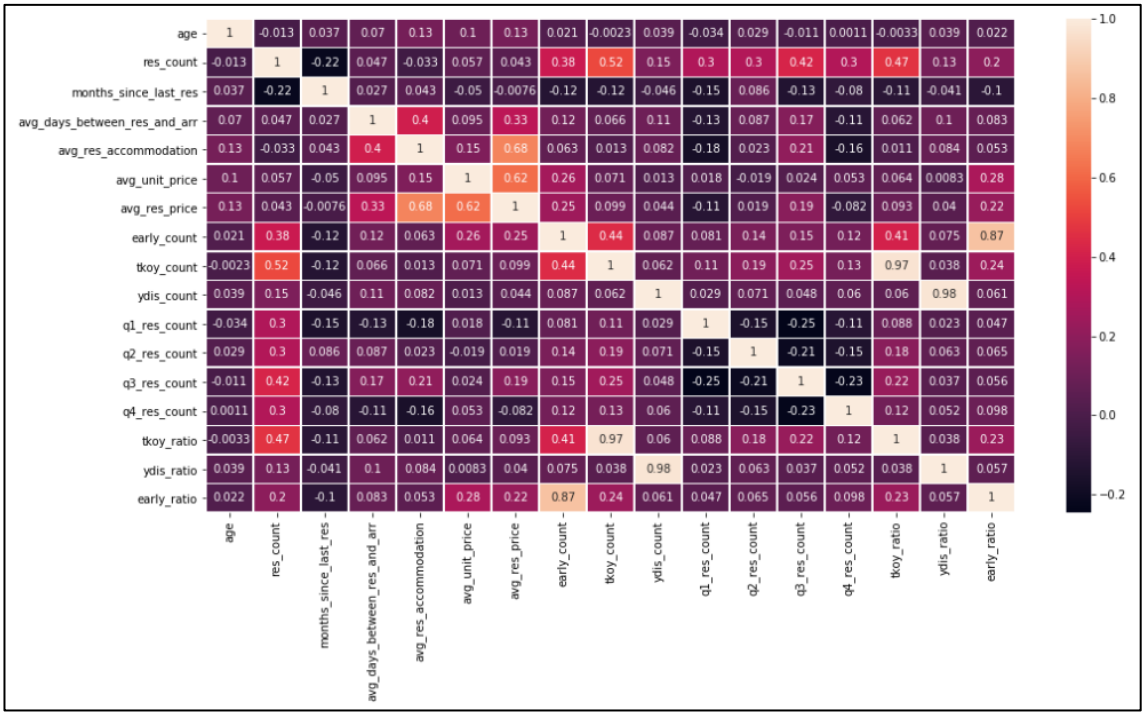


Birbirinden ayrı olmasını beklediğimiz kümelerimiz sarı, lacivert ve kırmızı renkler ile temsil edilmiştir. Bu durumda ilgili renklere sahip olan kayıtları veri setimizden okuyarak korelasyon grafiğinde hangi alanların birbiri ile ilişkili olduğunu gözlemleyerek, kümeleme işleminin hangi özellikleri baz alarak bu işlemi tamamladığını bulabiliriz. Şekil 3.12'te, Şekil 3.13'de ve Şekil 3.14'de kümelerin korelasyon grafiğini görebiliriz. Korelasyon işlemi bize veri içerisindeki sütunlar arası eksi veya artı yönlü bağlantıları bulmamızı sağlayan bir yöntemdir. Genellikle ısı haritaları üzerinde gösterilen grafiklerde eksi yönlü değerler, bir alandaki değer artarken diğer alandaki değerin azaldığını gösterir. Grafiğin en sağındaki renk çubuğunun en üst ve en altındaki değerler bağlantılarının güçlülük oranlarıdır. 1 ve -0,6 en yüksek ilişkiyi temsil ederken çubuğun ortasına doğru gidildiğinde değerler arası ilişkinin normal düzeyde olduğunu söyleyebiliriz.

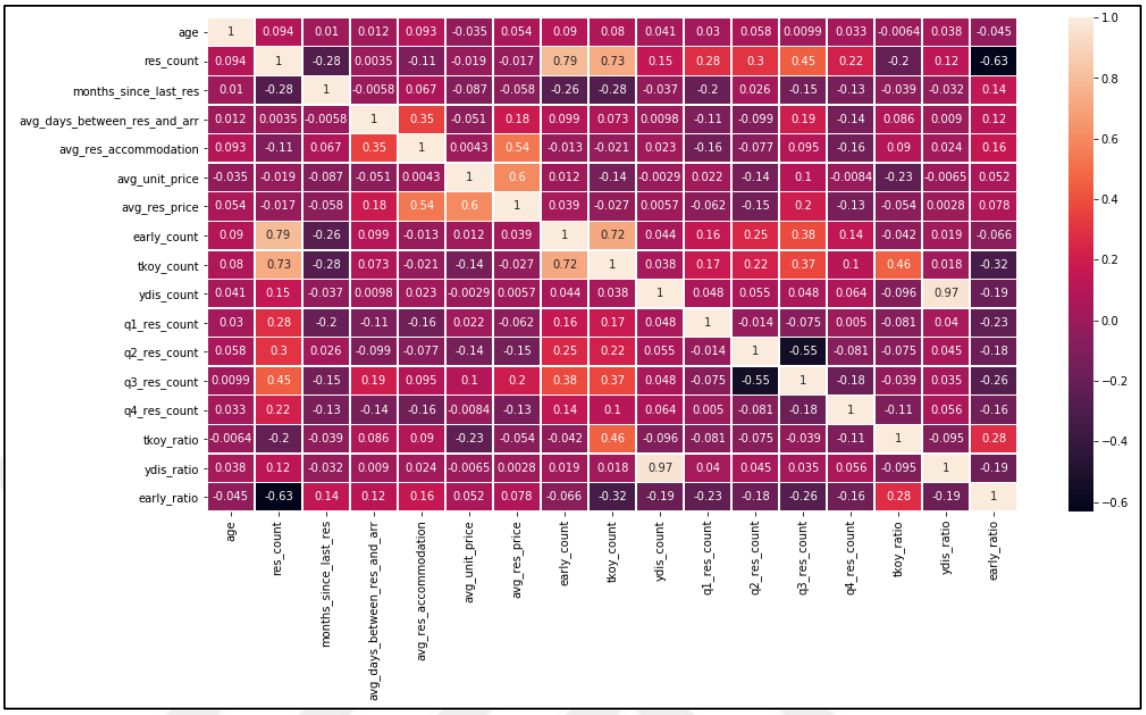
Şekil 3.12: Sarı renkli kümenin korelasyon grafiği



Şekil 3.13: Kırmızı renkli kümenin korelasyon grafiği



Şekil 3.14: Lacivert renkli kümenin korelasyon grafiği



4. BULGULAR

Aykırı veriler temizlendikten ve normalizasyon uygulandıktan sonra elimizde kalan 175144 satır veri üzerinde yapmış olduğumuz analizlere ait yorumlarda bulunarak, veri setimiz üzerinde yapacağımız yorumların doğru sonuç verip vermediğini ispatlamak için filtreler ve gruplamalar uygulayalım. Şekil 3.15'de ve Şekil 3.16'de korelasyona sokmuş olduğumuz alanlarımızın grup bazlı değerleri verilmiş ve diğerlerinden aykırı şekilde büyük veya küçük olan kayıtlar işaretlenmiştir. Bu işaretli alanlar "*k-means"* algoritmasının hangi alanları baz alarak kümeleme yaptığını bulma konusunda bize yardımcı olacaktır. Örneğin; "*early_ratio"* isimli alana bakacak olursak, bu alanın ortalama değer üzerinde kalan kayıtlarının lacivert kümede yoğunlaştığını söyleyebiliriz veya aynı şekilde *tkoy_ratio* alanına bakacak olursak kırmızı kümede ortalamanın üzerinde kalan çok az kayıt olduğunu söyleyebiliriz. Bu şekilde kümelerin karakteristik özelliklerini belirlemek de kolay olacaktır. Örneğin; kırmızı kümedeki misafirlerimiz az sayıda tatil köyü rezervasyonu yaptırıyor ancak "*q4_res_count"* alanına bakarak tatil

planlarının yılın son çeyreğinde yoğunlaştığını söyleyebiliriz. Şimdi buradan ve korelasyon grafiklerinden bazı çıkarımlar yaparak filtrelerimizi uygulayıp sonuçları görelim.

Şekil 3.15: Alanların, bulundukları küme içerisinde ortalamanın üzerinde kalan kayıtları yüzdesi

age	Kırmızı Küme: %45.61	Lacivert Küme: %43.51	Sarı Küme: %43.19
res_count	Kırmızı Küme: %24.65	Lacivert Küme: %24.85	Sarı Küme: %25.33
months_since_last_res	Kırmızı Küme: %51.94	Lacivert Küme: %52.99	Sarı Küme: %49.97
avg_days_between_res_and_arr	Kırmızı Küme: %26.37	Lacivert Küme: %47.79	Sarı Küme: %31.68
avg_res_accommodation	Kırmızı Küme: %34.88	Lacivert Küme: %44.19	Sarı Küme: %38.81
avg_unit_price	Kırmızı Küme: %35.68	Lacivert Küme: %41.73	Sarı Küme: %40.08
avg_res_price	Kırmızı Küme: %32.33	Lacivert Küme: %42.13	Sarı Küme: %38.61
early_count	Kırmızı Küme: %4.83	Lacivert Küme: %16.03	Sarı Küme: %9.81
tkoy_count	Kırmızı Küme: %4.09	Lacivert Küme: %17.89	Sarı Küme: %15.16
ydis_count	Kırmızı Küme: %1.12	Lacivert Küme: %0.51	Sarı Küme: %0.77
q1_res_count	Kırmızı Küme: %26.24	Lacivert Küme: %4.75	Sarı Küme: %33.25
q2_res_count	Kırmızı Küme: %25.76	Lacivert Küme: %39.93	Sarı Küme: %16.17
q3_res_count	Kırmızı Küme: %40.09	Lacivert Küme: %60.52	Sarı Küme: %48.96
q4_res_count	Kırmızı Küme: %25.34	Lacivert Küme: %8.33	Sarı Küme: %20.88
tkoy_ratio	Kırmızı Küme: %4.09	Lacivert Küme: %84.71	Sarı Küme: %85.42
ydis_ratio	Kırmızı Küme: %1.12	Lacivert Küme: %0.51	Sarı Küme: %0.77
early_ratio	Kırmızı Küme: %4.83	Lacivert Küme: %87.49	Sarı Küme: %9.81

Şekil 3.16: Kümelerdeki değerlerin alan bazlı ortalaması

	Alan	Genel Ort.	Kırmızı Küme Ort.	Lacivert Küme Ort.	Sarı Küme Ort.
0	age	0.491917	0.484590	0.490859	0.506132
1	res_count	0.056931	0.061824	0.052630	0.056832
2	months_since_last_res	0.355976	0.359189	0.353284	0.355661
3	avg_days_between_res_and_arr	0.263737	0.152571	0.423643	0.149347
4	avg_res_accommodation	0.366060	0.249272	0.477930	0.351171
5	avg_unit_price	0.299020	0.195343	0.372437	0.334380
6	avg_res_price	0.213207	0.080051	0.331247	0.214068
7	early_count	0.175648	0.016830	0.392899	0.033236
8	tkoy_count	0.248582	0.013706	0.381145	0.392020
9	ydis_count	0.007919	0.011208	0.005119	0.007682
10	q1_res_count	0.106621	0.149391	0.025379	0.187623
11	q2_res_count	0.166412	0.146857	0.226046	0.087187
12	q3_res_count	0.200580	0.163210	0.241271	0.186633
13	q4_res_count	0.096892	0.144408	0.044560	0.115735
14	tkoy_ratio	0.578865	0.012982	0.894201	0.932036
15	ydis_ratio	0.006726	0.009645	0.004216	0.006562

4.1 SARI RENKLİ KÜME

Şekil 3.12'de sarı renkli kümenin korelasyon haritasına bakarak "*res_count*" (toplam rezervasyon sayısı) ve "*tkoy_count*" (toplam tatil köyü rezervasyon sayısı) arasında artı(+) yönlü ilişki olduğunu söyleyebiliriz. Bu durumda toplam rezervasyon sayısı artarken tatil köyü rezervasyonlarının da arttığını görüyoruz. Bu kümede yer alan insanların davranışlarına bakıldığında daha fazla rezervasyon yaptıklarında tatillerinin çoğunu kültür turu veya yurtdışı seyahatlerinde geçirmek yerine tatil köylerinde geçirdikleri ortaya çıkıyor. Aynı şekle bakarak *"avg_res_price"* ile *"q3_res_count"* arasındaki pozitif yönlü ilişkinin de bize bu gruptaki misafirlerin bir kısmının yılın üçüncü çeyreğinde yaptıkları tatillerde diğer çeyreklere oranla daha fazla para harcadıklarını söylüyor. Aynı şekil "*res_count*" (toplam rezervasyon sayısı) ile "*avg_res_accommodation*" (rezervasyon başına geçirilen gece sayısı) arasında ise eksi(-) yönlü bir ilişkiyi de gösteriyor bize. Rezervasyon sayısı arttıkça tatillerdeki gece sayısının azaldığı anlamına geliyor bu. Türkiye'deki çalışanların yıllık izin günleri ve bu izinlerini tatil yaparak geçirdiklerini ve izinlerinin de limitli olduğu düşünüldüğünde çok fazla rezervasyon ancak kısa kısa tatiller yaptıklarını tahmin etmek çok da zor değil.

Tatillerinin çoğunu tatil köylerinde geçirenlerin sayısı yüzde 15.16 ile 5841 kişi olurken, yılın üçüncü çeyreğinde rezervasyon başına ortalama üzeri ücret ödeyenlerin sayısı ise yüzde 25.93 ile 9992 kişi. Tatillerini kısa tutarak çok sayıda rezervasyon yapmak isteyenlerin sayısı ise yüzde 18.28 ile 7043 kişi olarak karşımıza çıkıyor.

4.2 LACİVERT KÜME

Şekil 3.13'deki lacivert renkli kümenin grafiğinde, sarı renkli kümenin grafiğinden farklı olarak örnek gösterebileceğimiz ilişki "*early_count*" ve "*q3_res_count*" (yılın 3. çeyreğindeki tatiller) artı(+) yönlü bir ilişki olduğunu görüyoruz. Yani bu gruptaki erken rezervasyon dönemindeki indirimlerden faydalanıp tatile çıkan misafirler, tatillerini yılın

üçüncü çeyreğinde geçiriyor. Bu misafirlerin sayısı yüzde 11.74 ile 8486. Aynı şekil bize *"res_count"* ile *"q3_res_count"* arasında da pozitif yönlü ilişki olduğunu söylüyor. Yılın üçüncü çeyreğinde tatil yapmayı seven insanların rezervasyon sayılarının fazla olduğunu dolayısı ile de bu misafirlerin rezervasyonlarının büyük bir bölümünü de erken rezervasyon döneminde yaptığını söyleyebiliriz. Rezervasyon sayısı ile üçüncü çeyrekte tatil yapanların sayısı ise yüzde 18.18 ile 13138 kişi.

4.3 KIRMIZI RENKLİ KÜME

Şekil 3.14'de, kırmızı renkli kümedeki verilerin ortak özelliklerini incelediğimizde sarı ve lacivert kümedeki bulgulardan farklı olarak "*age*"(misafir yaşı) ile "*q1_res_count*"(yılın ilk çeyreğinde yapılan tatil) arasında eksi(-) yönlü ilişkinin olduğunu söyleyebiliriz. Yani kırmızı kümedeki misafirlerin yaşları ilerledikçe yılın ilk çeyreğinde tatil yapmaktan kaçınıyorlar ya da diğer bir deyişle, misafirlerin yaşları azaldıkça ilk kış tatili sayısı artıyor diyebiliriz. Yani gençler kış tatillerine daha çok gidiyor. Bu doğrultuda toplamda 64328 kaydın bulunduğu kırmızı küme içerisinde yaşı grup ortalamasının üzerinde kalan ve ilk çeyrekte tatil yapma sayıları ortalamanın altında kalan misafirlerimizin toplam sayısı yüzde 33.95 ile 21840 oluyor. Tüm veri setine oranı ise yüzde 12.47. Ayrıca yine bu küme içerisinde *"res_count"* arttıkça *"ydis_count"* sayılarının da arttığını görebiliriz. Bu artışa ait rakamlar her ne kadar küçük de olsa diğer gruplar ile karşılaştırıldığında fark büyük oluyor. Bu gruba ait sayı ise yüzde 1.12 ile 771 kişi. Mavi grupta 370 kişi bulunuyorken sarı grupta ise 296 kişi bulunuyor.

5. TARTIŞMA VE SONUÇ

Bu çalışmanın amacı, hali hazırda faaliyet gösteren bir turizm firmasının geçmiş verilerinin müşteri alışkanlıkları yönünden incelenerek, müşteri alışkanlıklarını ve profillerini ortaya çıkartmaktı. Bu çalışma için toplamda 200000 satır ve 18 sütundan oluşan müşteri bilgileri analiz edilmek üzere *python* programlama diline yüklendi. İlk adımda, veri setindeki boş değerlerin tespit edilmesi, eğer varsa yerlerine sütun ortalaması değerlerinin atanması işlemleri uygulandı. Daha sonra aykırı veriler veri setinden çıkartılarak temizlik aşaması tamamlandı. z-değerler ve en küçük-en büyük yöntemleri kullanılarak veriler normalize edildi. *Elbow* yöntemin sonucunda optimal küme sayısı belirlenerek k-means algoritması ile veri setimizdeki veriler 3 kümeye ayrıldı ve her küme için korelasyon çıktısı ısı haritası üzerinde gösterildi. Harita üzerinde analiz ve yorumlar yapıldı ve bu yorumların ispatları tezin bulgular kısmında gösterildi. Bu çalışma sonucunda turizm firması veri ambarında bulunan ve işlem görmemiş bilgiler anlamlı hale getirilerek kendi müşterilerinin profilleri belirlendi. Bu profillere göre satış ve pazarlama stratejileri belirlenecek ve müşteri alışkanlıklarına göre ürün ve hizmetler teklif edilecek. Satış faaliyetlerinde önemli rol oynayan bu yöntem birçok firmanın veri ambarındaki işlenmemiş verileri kullanarak satış faaliyetlerinin olumlu yönde artmasını sağlayabilir.

KAYNAKÇA

Süreli Yayınlar

[1] Dr. Yılmaz Argüden & Burak Erşahin, “Veri madenciliği & Veriden Bilgiye, Masraftan Değere” kitabı, 1. Basım, 2008, s. 5.

[3] S. Sevim Ve B. Aydın Sevim, "İzleyicinin Nabzını Tutmak: Büyük Veri, Tavsiye Algoritmaları ve *Netflix*," 3. Uluslararası İletişim Edebiyat Müzik ve Sanat Çalışmalarında Güncel Yaklaşımlar Kongresi, İstanbul, Türkiye, 2018

[4] Duru, N. ve Canbay, M., “Veri Madenciliği ile Deprem Verilerinin Analizi”, International Earthquake Symposium, Kocaeli, ss. 556-560, 2007.

[6] Türofed Turizm Bülteni, Bölüm 2.1, 2019 Mart, s. 18.

[8] Gregory Piatetsky-Shapiro, Usama Fayyad, Padhraic Smyth, “Knowledge Discovery and Data Mining: Towards a Unifying Framework”, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, 1996.

[9] Şule Özmen, İstanbul Ticaret Üniversitesindeki Veri Madenciliği ve Uygulama Alanları Konferansından, 2001.

[10] Ercan Çiçek. Müşteri İlişkileri Yönetimini Uygulama Sürecinde Başarıyı Etkileyen Faktörler, İktisadi ve İdari Bilimler Dergisi, Sayı:2, Cilt:5, 2005.

[11] Gordon Linoff & Michael J.A. Berry, “The Art and Science of Customer Relationship Management”, Amerika Birleşik Devletleri, Wiley Bilgisayar Yayıncılığı, 2000, ss.7-8.

[12] Dolf Zantinge & Pieter Adrians, “Data Mining” isimli kitap, İngiltere, 1997, s.5.

[13] Ethem Alpaydın, “Ham Veriden Altın Bilgiye Ulaşma Yöntemleri”, Bilişim Eğitim Semineri, 2000.

[15] Tong Wu, Jothishankar, M. C, Jiun-Yan Shıau, Johnie Roberts “Appling Data Mining to Defect Diagnosis”, “Advanced Manufacturing Systems” isimli dergi, baskı 3, 2004, s. 71.

[17] ODABAŞI Yavuz- Gülfidan BARIŞ, Tüketici Davranışları, İstanbul 2003.

[18] Vahaplar, A., & İnceoğlu, M. Veri madenciliği ve elektronik ticaret. Türkiye’de İnternet Konferansları, İstanbul, 2001 (Kasım)

[19] Shen, L., Dickerson, K., Hawley, J., “E-commerce adoption for supply chain management in US apparel manufacturers”. “Textile and Apparel” isimli dergi, 2004, ss.1-11.

[22] Feng Y., Huirnin Q., & Hemin J. E-ticarette değiştirilmiş ID3 algoritmasına dayalı olarak müşteri grupları için veri madenciliği uygulaması üzerine çalışma. Uluslararası Bilgisayar Bilimi ve Bilgi İşlem Konferansı. Şangay, Çin, 2012 Ağustos.

[23] Huang W., & Zhou X. E-ticarette küme tabanlı veri madenciliği tekniklerinin araştırılması, Yönetim ve Hizmet Bilimi Konferansı, Wuhan, Çin, 2009, Eylül.

[24] Cheng, Y. & Xiong, Y. E-ticarette veri madenciliği teknolojisinin uygulanması. Uluslararası Bilgisayar Bilimi-Teknolojisi ve Uygulamaları Forumu, Washington, USA, 2009, Aralık

[27] Hand & D.J. “Data Mining: Statistics and More”, “The American Statistician” Dergisi, Cilt 52, 1998, 112 - 118.

[28] Kitler R. ve Wang W. “The Emerging Role of Data Mining”, Solid State Teknoloji Dergisi, 1998, Cilt 42, No 11, 45.

[29] Minsky, M. & Papert, S, Perseptronlar. MIT Press Gazetesi, Cambridge, 1969.

[31] D. Hawkins. “Aykırı Değerlerin Belirlenmesi” isimli kitap, 1980.

[25] Jacobs, P. “Data Mining: What General Managers Need to Know”, Harvard Yönetimi Güncellemesi, 1999, Cilt 4, No 10, s. 8.

[26] Şengül Doğan ve İbrahim Türkoğlu, "Hypothyroidi and Hyperthyroidi Detection from Thyroid Hormone Parameters by Using Decision Trees", Doğu Anadolu Bölgesi Araştırmaları Dergisi, 2007 Ocak, Cilt 5, No 2, s. 163 - 169.

[33] Jiawei Han, Illinois Üniversitesi, “Veri Madenciliği Teknikleri isimli” kitap, baskı 3, 2012.

[35] Zhang, T. R. Ramakrishnan ve M. Livny. *BIRCH*: Büyük veritabanları için verimli bir kümeleme yöntemi. ACM SIGMOD uluslararası veri yönetimi konferansı, Montreal, Quebec, Kanada, s.107. 1996.

[36] Tong, H. ve U. Kang. Big Data Clustering. C. C. Aggarwal ve C. K. Reddy (Ed.). Data Clustering: Algorithms and Applications içinde. Florida: CRC Baskısı, ss. 259-276, 2014.

Diğer Yayınlar

[2] Ankara Yıldırım Beyazıt Üniversitesi web sitesi, Veri Madenciliği Araştırma Alanları, 2017.

[5] Gümüş, H. Türkiye'de ulusal turizm örgütlerinin yapısal analizi ve turizm pazarlamasına katkılarına yönelik bir araştırma çalışması. Yayımlanmamış yüksek lisans tezi, Balıkesir Üniversitesi, Balıkesir, 2008.

[7] Mehmet Berkay Can, Eren Çamur, Mine Koru, Ömer Özkan & Zeynep Rzayeva, Veri Kümelerinden Bilgi Keşfi: Veri Madenciliği, Başkent Üniversitesi, 2013 Ağustos.

[14] Bünchner, Alex G. Sarabjot S. Anand & John G. Hughes. "Data Mining in Manufacturing Environments: Goals, Techniques and Applications", isimli çalışması, 1997, s. 1.

[16] Erdem Uzun, UzakRota, Teknolojinin Turizm Şirketleri Üzerindeki Etkisi, 2018 Ocak, www.uzakrota.com

[20] Kalikov, A. "Veri madenciliği ve bir e-ticaret uygulaması" isimli yüksek lisans tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 2006.

[21] Yuantao J. S. Y. Hedef müşteri davranışını analiz etmek için e-ticaret verilerini incelemek, Bilgi Keşfi ve Veri Madenciliği, Las Vegas, 2008, Ocak.

[30] Serkan Savaş, Nurettin Topaloğlu & Mithat Yılmaz, Veri madenciliği ve Türkiye'deki Uygulama Örnekleri, 2012, s. 2.

[32] Hong Hai Do & Erhard Rahm, "Data Cleaning: Problems and Current Approaches", Leipzig Üniversitesi, Almanya, 2000.

[34] Yard. Doç. Dr. ULAŞ AKKÜÇÜK, Boğaziçi Üniversitesi, Veri Madenciliği: Kümeleme ve Sınıflandırma Algoritmaları, 2011.