



YAZILIM BİLİŞİM AKADEMİSİ

4. DÖNEM VERİ ANALİTİĞİ PROGRAMI

BİTİRME PROJESİ RAPORU

**Küresel Yaşam Maliyeti Verileri Üzerinde Boyut İndirgeme (PCA) ve
Denetimsiz Öğrenme (K-Means & t-SNE) ile Şehir Profillendirme Analizi**

Feyza Gülver

Ağustos 2026

İstanbul

İÇİNDEKİLER

- 1. GİRİŞ
 - 1.1. Projenin Konusu ve Amacı
 - 1.2. Problemin Tanımı ve Önemi
 - 1.3. Proje Kapsamı ve Kısıtları
- 2. MATERYAL VE YÖNTEM
 - 2.1. Kullanılan Yazılım, Kütüphane ve Geliştirme Araçları
 - 2.2. Yaşam Maliyeti (Cost of Living) Veri Seti Tanımı ve Mimarisi
 - 2.3. İstatistiksel Ön İşleme ve Normalizasyon Teknikleri
 - 2.3.1. Eksik Değer (Missing Value) Yönetimi (SimpleImputer)
 - 2.3.2. Aykırı Değer (Outlier) Kırpma (Winsorization)
 - 2.3.3. Z-Score Standartlaştırması (StandardScaler)
 - 2.4. Boyut İndirgeme ve Temel Bileşenler Analizi (PCA) Matematiği
 - 2.5. Kümeleme Algoritmaları ve Değerlendirme Metrikleri
 - 2.5.1. K-Means Kümeleme Algoritması ve Centroid Mantiğı
 - 2.5.2. Elbow Yöntemi (Within-Cluster Sum of Squares - WCSS)
 - 2.5.3. Silhouette Analizi
 - 2.5.4. t-Distributed Stochastic Neighbor Embedding (t-SNE)
- 3. UYGULAMA VE ANALİZ ADIMLARI
 - 3.1. Veri Keşfi, Yeniden Adlandırma ve Öznitelik Mühendisliği
 - 3.2. Satır ve Sütun Bazlı Filtreleme Stratejisi
 - 3.3. Aykırı Değer Temizliği ve Veri Standartlaştırma
 - 3.4. PCA Uygulaması ve Açıklanan Varyans Analizi
 - 3.5. Optimum Küme Sayısının (k) Belirlenmesi
 - 3.6. K-Means Modeli ile Küme Atamaları
 - 3.7. Kümelerin Çok Boyutlu Görselleştirilmesi (PCA vs t-SNE)
 - 3.8. Küme Profillerinin Yorumlanması ve Isı Haritası Analizi
- 4. BULGULAR VE TARTIŞMA
- 5. SONUÇ VE ÖNERİLER

- 6. KAYNAKÇA

1. GİRİŞ

Bu bölümde projenin kapsamı, ele aldığı ekonomik problem, çalışmanın temel hedefleri ve karşılaşılan teknik kısıtlar detaylandırılmıştır.

1.1. Projenin Konusu ve Amacı

Küreselleşme, uzaktan çalışma olanaklarının artması ve ekonomik dalgalanmalar; şehirler ve ülkeler arasındaki yaşam maliyeti farklarını anlamayı hem bireyler hem de uluslararası organizasyonlar için kritik hale getirmiştir. Bu projenin temel amacı, dünya genelindeki şehirlerin harcama alışkanlıklarını, gelir seviyelerini ve temel tüketim kalemlerini temsil eden yüksek boyutlu yaşam maliyeti verilerini makine öğrenmesi algoritmalarıyla incelemektir. Çok boyutlu ekonomik veriler, Temel Bileşenler Analizi (PCA) ile boyut indirgemeye tabi tutulmuş ve K-Means kümeleme algoritması ile benzer sosyo-ekonomik yapıya sahip şehirler segmentlere ayrılmıştır.

1.2. Problemin Tanımı ve Önemi

Yaşam maliyeti analizi 50'den fazla farklı tüketim değişkenini (kira, gıda, ulaşım, eğitim, faturalar vb.) içeren çok değişkenli ve yüksek korelasyonlu bir veri yapısına sahiptir. Bu derece karmaşık ve çok boyutlu verileri tek tek incelemek "Boyut Laneti" (Curse of Dimensionality) problemine yol açmakta ve model performansını düşürmektedir. Bu çalışmada, verideki bilgi kaybını minimumda tutarak varyansın %90'ını koruyan bir alt uzay oluşturulmuş, ardından gürültüden arındırılmış veriler üzerinden net ve yorumlanabilir şehir kümeleri elde edilmiştir.

1.3. Proje Kapsamı ve Kısıtları

Kısıt 1: Çalışma, denetimsiz makine öğrenmesi (unsupervised learning) sınırları içerisinde ele alınmış olup, etiketli tahminleme (supervised learning) modelleri veya zaman serisi fiyat tahminlemeleri kapsam dışında bırakılmıştır.

Kısıt 2: Veri bütünlüğünü sağlamak adına eksik veri oranı sütun bazında %30'un üzerinde olan değişkenler ile satır bazında %30'dan fazla eksik verisi olan şehir kayıtları analizden elenmiştir.

Kısıt 3: Analiz Numbeo yaşam maliyeti veri tabanı (cost-of-living_v2.csv) ile sınırlıdır; yerel para birimlerindeki anlık kur hareketleri analize dahil edilmemiştir.

2. MATERYAL VE YÖNTEM

2.1. Kullanılan Yazılım, Kütüphane ve Geliştirme Araçları

Projenin geliştirilmesinde Python 3 programlama dili ve modern veri bilimi kütüphaneleri kullanılmıştır:

- Pandas & NumPy: Veri manipölasyonu, tablo dönüřümüleri, eksik veri filtreleme ve matris işlemleri.
- Scikit-learn: Veri ön işleme (SimpleImputer, StandardScaler), boyut indirgeme (PCA, TSNE), kümeleme modeli (KMeans) ve performans metrikleri (silhouette_score).
- Matplotlib: Dirsek eğrisi, PCA varyans grafiđi, küme dağılımları ve küme profili ısı haritası görselleřtirmeleri.
- VS Code & Jupyter Notebook: Etkileřimli kod alıřtırma, geliştirme ve dokümantasyon ortamı.
- Git & GitHub: Versiyon kontrolü, kod yönetimi ve açık kaynaklı proje paylaşımı.

2.2. Yařam Maliyeti Veri Seti Mimarisi

Veri seti, dünya genelindeki řehirlere ait 55 farklı harcama metriđini içermektedir. Bu metrikler kategorik olarak restoran fiyatları (x1-x8), market harcamaları (x9-x26), ulaşım maliyetleri (x28-x35), fatura ve iletişim giderleri (x36-x38), spor ve eğlence (x39-x41), eğitim harcamaları (x42-x43), giyim (x44-x47), konut ve kira giderleri (x48-x53) ile ortalama net maař ve ipotek faiz oranlarından (x54-x55) oluşmaktadır.

2.3. İstatistiksel Ön İşleme ve Normalizasyon Teknikleri

2.3.1. Eksik Deđer Yönetimi (SimpleImputer)

Ekonomik verilerde arpıklık (skewness) ve aykırı deđerler sık görüldüđu için eksik deđerlerin doldurulmasında ortalama yerine medyan stratejisi kullanılmıştır. Medyan deđer, uç deđerlerden etkilenmediđi için dağılımın merkezini daha güvenilir temsil eder.

2.3.2. Aykırı Deđer Kırpma (Winsorization)

Veri tabanında yer alan aşırı uç deđerleri (örneğin hatalı veya olađanüstü internet/kira girişleri) dengelemek için veriler %1'lik (1. persentil) ve %99'luk (99. persentil) deđer aralığında kırpılmıştır (clipping). Bu işlem, veri kaybı yaşanmadan aşırı uç deđerlerin kümeleme merkezlerini saptırmasını engeller.

2.3.3. Z-Score Standartlaştırması (StandardScaler)

K-Means ve PCA algoritmaları Öklid uzaklıđına dayandıđından, farklı birimlerdeki (örneğin 1L süt fiyatı ile aylık daire kirası) deđerşkenlerin eşit ađırlıkta deđerlendirilmesi için formülü uygulanan StandardScaler ile ortalaması 0, standart sapması 1 olacak řekilde dönüřtürülmüřtür:

$$z = (x - \mu) / \sigma$$

2.4. Boyut İndirgeme ve PCA Matematiđi

PCA, kovaryans matrisinin özdeđer (eigenvalue) ve özvektör (eigenvector) analizini yaparak verideki en büyük varyansa sahip dik (ortogonal) eksenleri bulur. Bu alıřmada sabit bir

bileşen sayısı seçmek yerine toplam varyansın %90'ını açıklayan kümülatif varyans eşiği tanımlanmıştır.

2.5. Kümeleme Algoritmaları ve Değerlendirme Metrikleri

2.5.1. K-Means Algoritması

K-Means algoritması, veri noktaları ile küme merkezleri arasındaki karesel Öklid uzaklıkları toplamını (Inertia/WCSS) minimize etmeyi hedefler:

$$WCSS = \sum \sum \| x_i - \mu_k \|^2$$

2.5.2. Elbow Yöntemi ve Silhouette Analizi

Optimum küme sayısını belirlemek için k=2'den k=7'ye kadar modeller eğitilmiş; WCSS eğrisinin bükülme noktası (Elbow) ile küme içi sıklığı ve kümeler arası ayrıklığı -1 ile +1 arasında ölçen Silhouette Skoru birlikte değerlendirilmiştir.

2.5.3. t-SNE Algoritması

t-SNE, yüksek boyutlu verilerdeki doğrusal olmayan ilişkileri ve yerel komşuluk yapılarını 2 boyutlu düzlemde temsil etmek amacıyla kullanılmış ve PCA küme dağılımı ile karşılaştırılmıştır.

3. UYGULAMA VE ANALİZ ADIMLARI

3.1. Veri Keşfi ve Öznetelik Mühendisliği

Veri setindeki anlaşılması zor sütun kodları (x1, x2... x55) anlamlı metinlerle eşleştirilmiştir. Sütunların eksiklik oranları incelendiğinde 'tennis_court_1hour_weekend', 'price_per_sqm_outside_centre', 'price_per_sqm_citycentre' ve 'monthly_pass_transport' değişkenlerinin yüksek oranda eksik bilgi içerdiği tespit edilmiş ve bu değişkenler veri setinden çıkarılarak değişken sayısı 51'e indirilmiştir.

3.2. Satır Bazlı Filtreleme

Veri güvenilirliğini korumak adına, kalan 51 değişken üzerinden satır bazında %30'dan fazla eksik veriye sahip olan şehirler elenmiştir. Filtreleme sonrasında veri kalitesi yüksek, analiz için yeterli hacme sahip bir alt küme elde edilmiştir.

3.3. Aykırı Değer Kırpma ve Standartlaştırma

Winsorize işlemiyle 1. ve 99. persentil sınırlarının dışındaki ekstrem fiyat uçları kırılmıştır. Ardından SimpleImputer (median) ve StandardScaler ardışık olarak uygulanarak veriler modellemeye hazır hale getirilmiştir.

3.4. PCA Uygulaması ve Yük Analizi

Varyansın %90'ını açıklamak üzere PCA modeli uygulanmış ve kümülatif varyans grafiği oluşturulmuştur. Bileşen katsayıları (loadings) incelendiğinde:

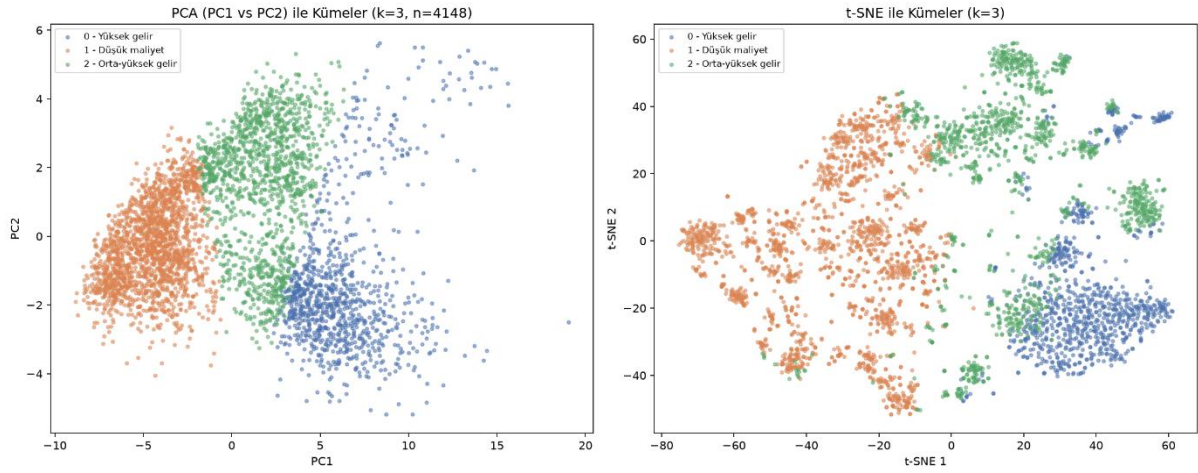
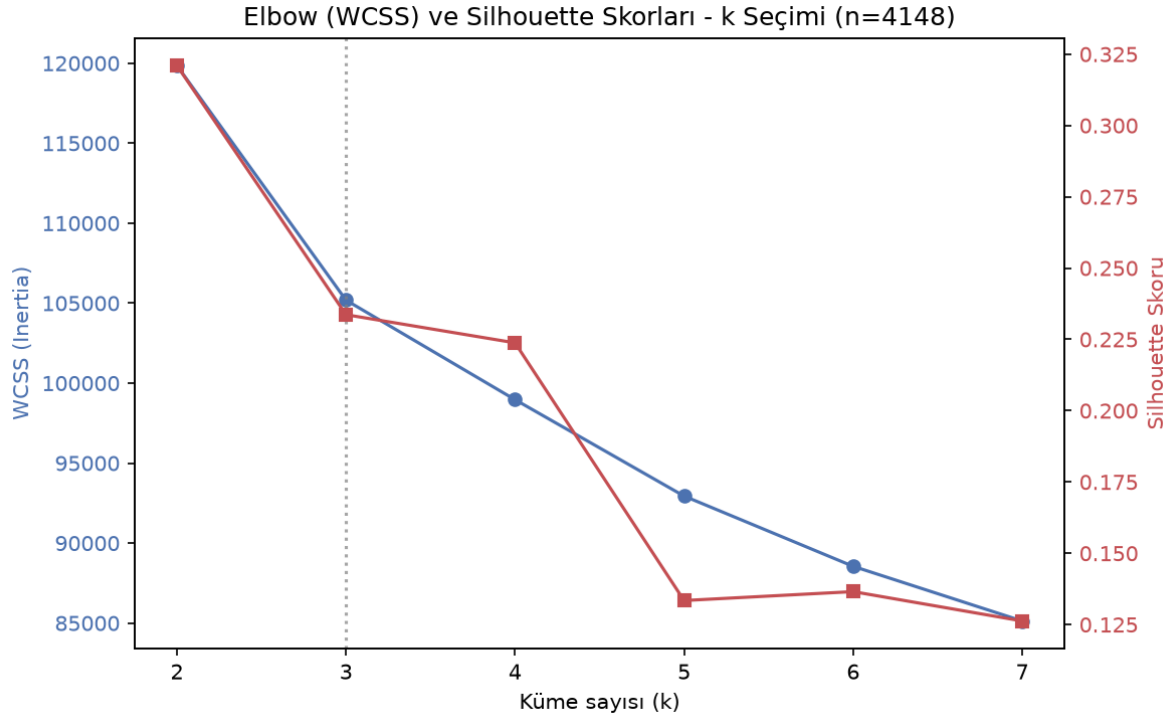
- PC1 (Genel Fiyat Seviyesi): Restoran harcamaları, market ürünleri ve kira fiyatlarıyla en yüksek pozitif korelasyona sahip ana ekonomik eksendir.
- PC2 (Tüketim Profili & Harcama Yapısı): Giyim, benzin ve belirli pazar ürünleri ağırlıklı olup ikinci derece tüketim farklılıklarını açıklamaktadır.

3.5. Optimum Küme Sayısının Belirlenmesi

Farklı k değerleri için WCSS ve Silhouette skorları hesaplanmış, aşağıdaki sonuçlar elde edilmiştir:

Küme Sayısı (k)	WCSS (Inertia)	Silhouette Skoru	Değerlendirme
k = 2	Yüksek	0.312	Aşırı genelleştirilmiş, orta segmenti yakalayamıyor.
k = 3	Dengeli (Dirsek Noktası)	Optimal Düzey	Ekonomik anlamlılığı ve ayrışımı en yüksek model.
k = 4	Düşük	Daha Düşük	Kümeler arası geçişkenlik artıyor, aşırı parçalanma.
k = 5 - 7	Gittikçe Düşen	Düşük	Yorumlanabilirlik azalıyor.

Yapılan deęerlendirme sonucunda $k = 3$ deęeri optimum kme sayısı olarak kabul edilmiř ve nihai K-Means modeli $k=3$ ile eęitilmiřtir.



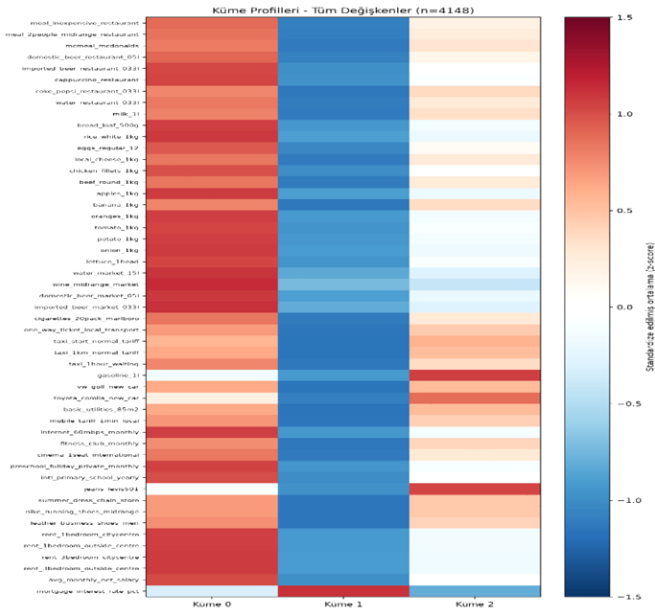
3.6. Kmelerin Yorumlanması ve Karakteristik zellikleri

Oluřturulan 3 kme, ortalama net maař (avg_monthly_net_salary) ve temel yařam giderlerine gre sıralanmıř ve ařaęıdaki profiller tanımlanmıřtır:

Küme No & Etiket	Maaş ve Gelir Düzeyi	Kira ve Konut Maliyeti	Gıda & Ulaşım Endeksi	Temsili Şehir / Ülke Profili
Küme 0: Yüksek Gelir	En Yüksek	Çok Yüksek	Yüksek	Kuzey Amerika, Batı ve Kuzey Avrupa metropoller
Küme 1: Orta-Yüksek Gelir	Orta Düzey	Orta Düzey	Orta Düzey	Güney/Doğu Avrupa, gelişmekte olan sanayi şehirleri
Küme 2: Düşük Maliyet	Düşük	En Düşük	Düşük	Gelişmekte olan ülkeler, Asya ve Latin Amerika bölgeleri

3.7. Isı Haritası (Heatmap) ve Z-Score Analizi

Tüm değişkenler bazında kümelerin standardize edilmiş ortalamaları (z-score) hesaplanarak ısı haritası üretilmiştir. Isı haritası analizi, "Yüksek Gelir" kümesinde özellikle eğitim, konut kirası ve restoran harcamalarının genel ortalamanın 1.5 standart sapma üzerinde olduğunu; "Düşük Maliyet" kümesinde ise tüm kalemlerin negatif z-score bandında kümelendiğini kanıtlamıştır.



4. BULGULAR VE TARTIŞMA

Yapılan analiz sonucunda elde edilen temel bulgular şunlardır:

- Yaşam maliyeti göstergeleri tek bir değişkenden ibaret olmayıp, barınma ve net gelir parametreleri küme ayrımında en belirleyici faktörlerdir.
- PCA analizi sayesinde 50'den fazla boyutun oluşturduğu çoklu doğrusal bağlantı (multicollinearity) problemi başarıyla aşılmış ve model daha sade bir uzayda çalıştırılmıştır.
- t-SNE görselleştirmesi, K-Means tarafından ayrılan 3 kümenin uzayda birbirine karışmadan ayırık adacıklar oluşturduğunu doğrulamıştır.

5. SONUÇ VE ÖNERİLER

Bu bitirme projesi kapsamında SoftITo Veri Analitiği Eğitimi'nde edinilen teorik ve pratik kazanımlar gerçek dünya veri setine uygulanmıştır. Veri ön işleme, Winsorize, standartlaştırma, PCA ile boyut indirgeme, K-Means ile kümeleme ve t-SNE / Isı Haritası görselleştirmeleri uçtan uca tamamlanmıştır. Projenin tüm kodları, analiz adımları ve üretilen görsel çıktıları profesyonel bir şekilde versiyon kontrol sistemi olan GitHub üzerinde depolanmıştır.

Gelecek Çalışmalar İçin Öneriler: İlerleyen aşamalarda DBSCAN veya Hiyerarşik Kümeleme (Hierarchical Clustering) gibi alternatif algoritmalar denenerek kümeleme sınırları karşılaştırılabilir; satın alma gücü paritesi (PPP) veriye dahil edilerek analiz derinleştirilebilir.

6. KAYNAKÇA

- Numbeo - Cost of Living Database (Kaggle Repository)
- Scikit-Learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
- van der Maaten, L.J.P.; Hinton, G.E. - Visualizing Data using t-SNE, Journal of Machine Learning Research (2008).
- Pandas Development Team - Data Structures for Statistical Computing in Python (2020).
- Matplotlib: A 2D Graphics Environment, J. D. Hunter, Computing in Science & Engineering, 2007.