# BANKA MÜŞTERİSİ KAYIP ANALİZİ

Recep Furkan Koçyiğit

furkan.kocyigit@std.yildiz.edu.tr

Bilgisayar Mühendisliği Bölümü

Elektrik Elektronik Fakültesi, Yıldız Teknik Üniversitesi

## Yöntem

Bu ödevde, bir alışveriş merkezine ait müşteri bilgilerine göre müşteri segmentasyonu yapılması hedeflenmektedir.

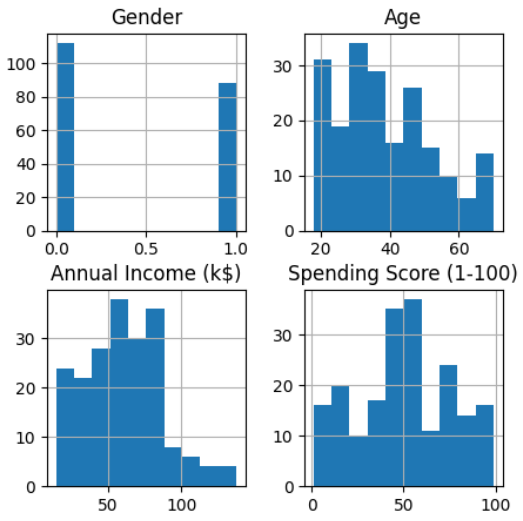
Bu kapsamda ilgili veri indirildikten sonra boş veri ve tekrar eden veri kontrolü yapılmış, id bilgisi her müşteri için benzersiz olduğundan kaldırılmıştır. Cinsiyet bilgisini tutan özellik nümerik olarak kodlanmıştır. Daha sonra veri standardize edilmiş, veri dağılımları ve korelasyon matrisleri incelenmiştir.

K-Means algoritması kullanılarak farklı k değerleri için segmentasyon yapılmış, en uygun k değerini bulmak için Elbow Yöntemi kullanılmıştır. En uygun k değeri ve en fazla hataları veren iki farklı k değerleri için verideki iki özelliğe bağlı olarak kümeleme işlem sonucu grafik haline getirilmiştir.

Ödev kapsamında kullanılan veri Mall Customer Segmentation Data | Kaggle adresinden elde edilmiştir. Bu veride 200 adet müşteri bilgisi bulunmaktadır.

## Veri Kümesindeki Özelliklere Ait Bilgiler

1. CustomerId: Müşteri numarası(Integer). Özellik olarak kullanılmamıştır.
2. Gender: Müşterinin cinsiyeti(String)
3. Age: Müşterinin yaşı(Integer)
4. Annual Income (k$): Müşterinin yıllık geliri (Integer).
5. Spending Score (1-100): Müşterinin harcama skoru(Integer).



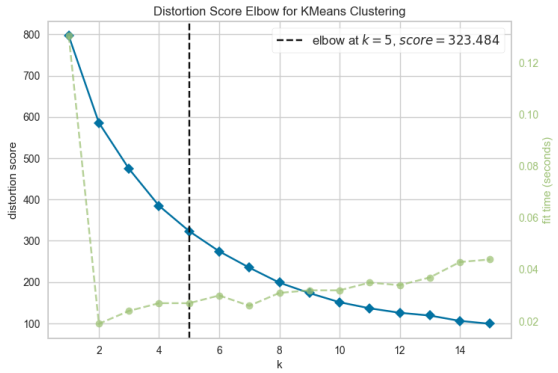

*Şekil 1 Verinin Dağılımı*

Veriye ait korelasyon matrisi aşağıdaki gibidir:


*Şekil 2 Korelasyon Matrisi*

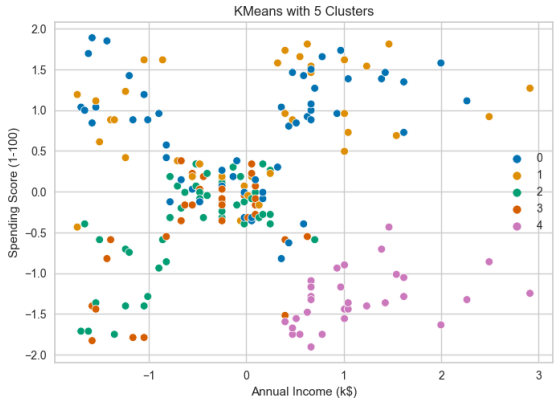
Korelasyon matrisine bakıldığında birbiri ile yüksek korelasyona sahip özellik olmadığı görülmüştür.

## Uygulama

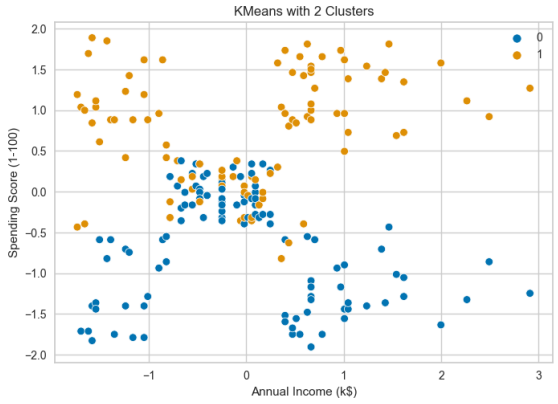



*Şekil 3 Elbow Yöntemi*

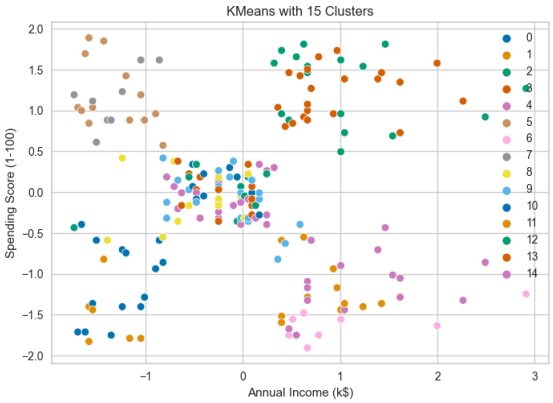
1 ile 15 arasındaki k değerleri için kümeleme işlemi yapıldığında en uygun k değerinin 5 olduğu saptanmıştır.




*Şekil 4 k=5 için kümeleme işlemi*




*Şekil 5 k=2 için kümeleme işlemi*




*Şekil 6 k=15 için kümeleme işlemi*

## Sonuç

Bu çalışmada müşteri segmentasyonu gibi kaç farklı küme olduğunu bilmediğimiz bir veri için en uygun küme sayısını bulmaya çalıştık. Elde ettiğimiz sonuçlarda en uygun küme sayısının 5 olduğu görüldü. K değerinin 2 seçildiğinde ise yetersiz küme sayısı olduğu için daha fazla hata meydana geldi. K değerini 15 seçildiğinde ise çok fazla küme sayısı olduğu için hatanın düşük gözüktüğü ama modelin yorumlanabilirliğini düşürdüğü gözlemlendi.