1. K-Means Əsas Anlayışlar 💡

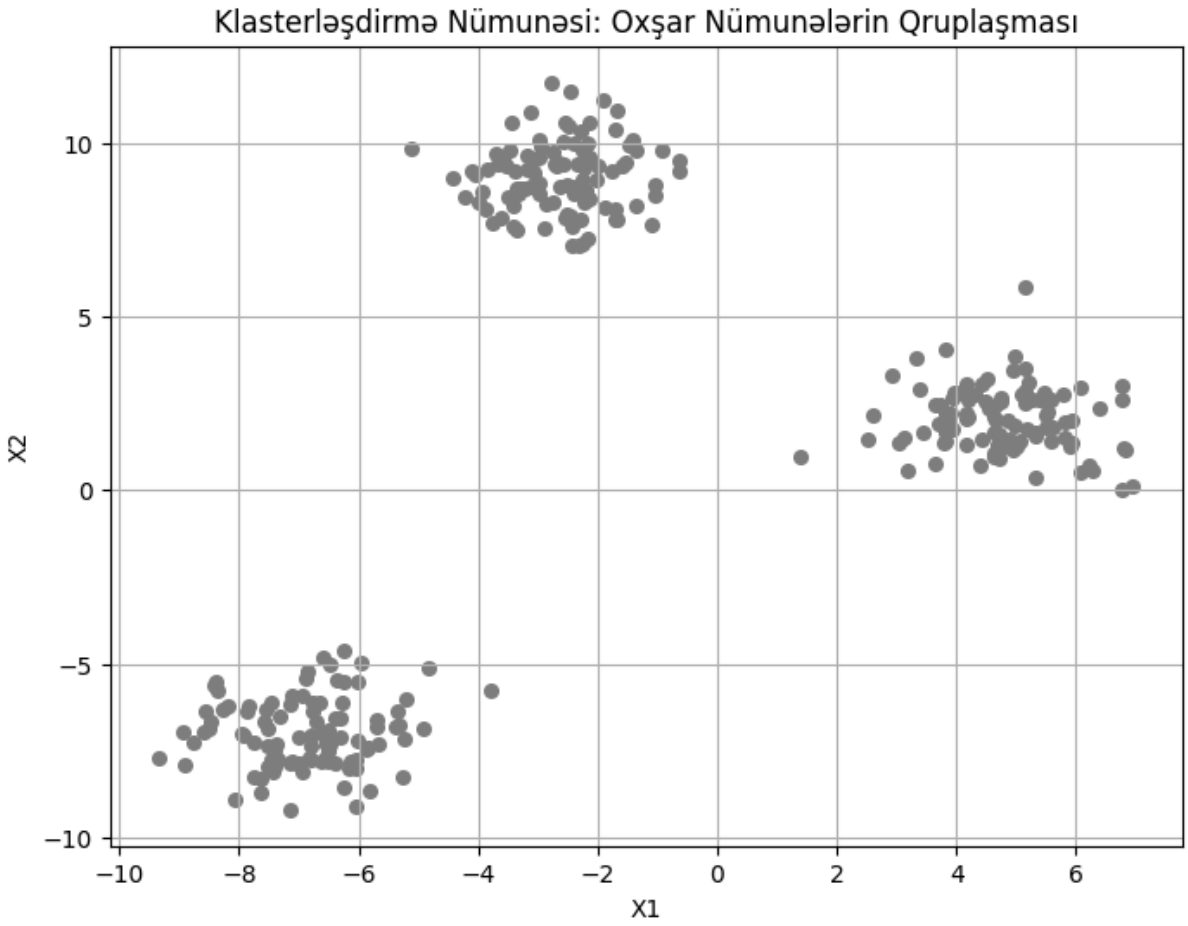
1.1. Klasterləşdirmə (Clustering) Nədir? 🔍

Klasterləşdirmə, süni intellektdə və maşın öyrənməsində **unsupervised learning** kateqoriyasına aid olan metodlardan biridir. Burada məqsəd verilənlər dəstində hər hansı əvvəlcədən təyin olunmuş etiket olmadan, yəni **label**siz, oxşar nümunələri bir qrupa, fərqli nümunələri isə digər qrupa ayırmaqdır. Məsələn, əgər əlində milyonlarla müştəri məlumatı varsa və bunlar üçün əvvəlcədən heç bir qrupa aidlik etiketi yoxdursa, klasterləşdirmə vasitəsilə müştəriləri davranışlarına, demoqrafik xüsusiyyətlərinə və ya satınalma vərdişlərinə görə fərqli seqmentlərə ayırmaq olar.

Bu yanaşma, verilənlərdə gizli qrupların, yəni **klasterlərin** tapılmasını təmin edir. Hər klaster daxilində nümunələr bir-birinə daha çox oxşayır, fərqli klasterlər isə mümkün qədər bir-birindən uzaq olur. Burada oxşarlıq çox vaxt məsafə və ya oxşarlıq ölçüsü ilə qiymətləndirilir. Klasterləşdirmə tətbiq sahələrinə görə müxtəlif ölçü və formalarda olur — bəzən klasterlər sıx və sferik ola bilər, bəzən isə qeyri-mütəşəkkil və dəyişkən formalı.

In [2]:

In [4]:



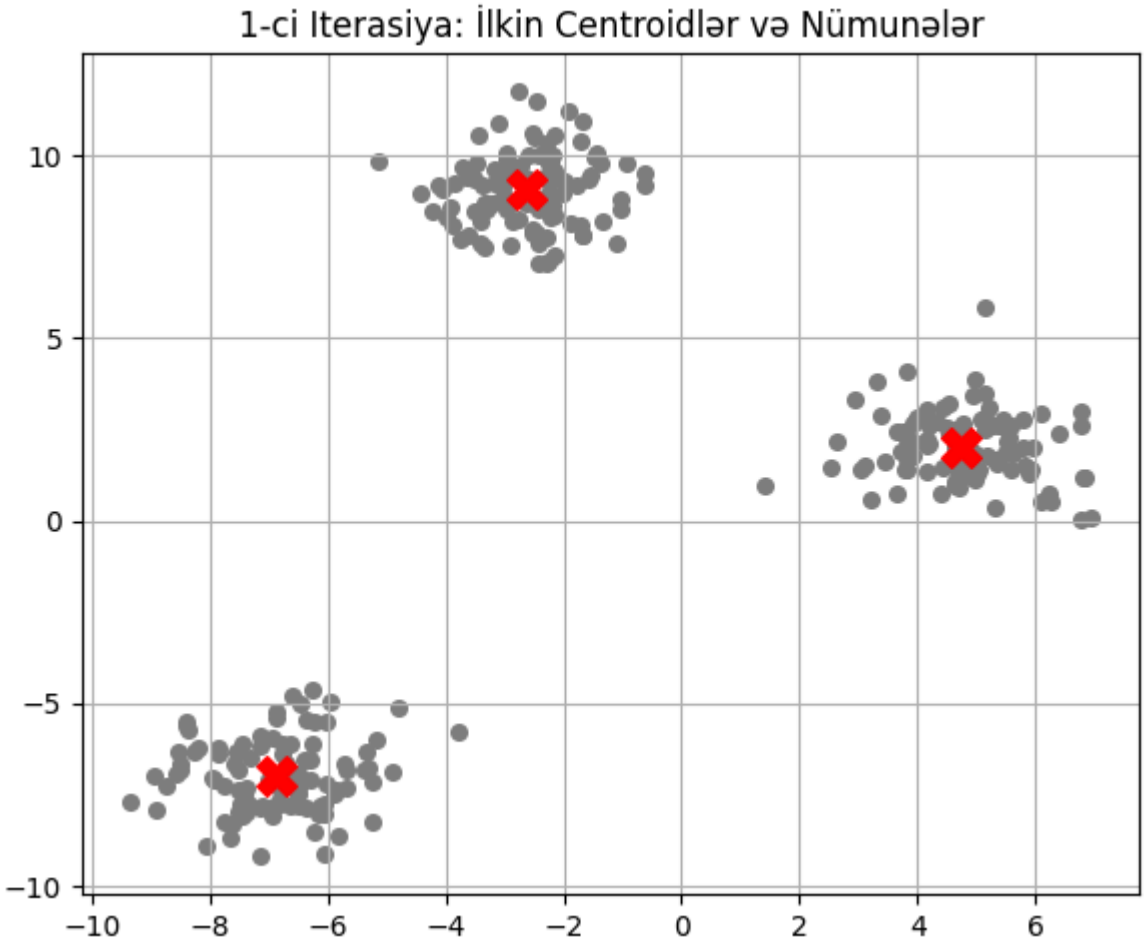
1.2. K-Means Alqoritmasının Tərifi ⚙️

K-Means, verilənləri K sayda klasterə bölən, çox sadə və geniş istifadə olunan klasterləşdirmə alqoritmidir. Burada K əvvəlcədən istifadəçi tərəfindən təyin olunmalıdır. Alqoritm aşağıdakı şəkildə işləyir:

- İlk olaraq, K ədəd klaster mərkəzi — **centroid** — təsadüfi və ya daha yaxşı başlanğıc strategiyası ilə seçilir.
- Hər bir nümunə özünə ən yaxın centroid-ə aid edilir. Bu, məsafə ölçüsü, çox vaxt **Euclidean distance** əsasında həyata keçirilir.
- Klaster təyinatları tamamlandıqdan sonra, hər klaster üçün yeni centroid hesablanır. Bu yeni centroid klasterə aid nümunələrin vektorlarının ortalamasıdır.
- Bu təyinat və yenilənmə addımları iterativ olaraq davam etdirilir. Hər iterasiya sonrası centroidlər yenilənir, nümunələr isə yenidən ən yaxın klasterə aid edilir.
- Proses klaster təyinatları dəyişmədikdə və ya maksimum iterasiya sayına çatdıqda dayanır.

Bu iterativ yanaşma ilə K-Means, verilənlərdəki strukturu ortaya çıxarır. Onun sadəliyi və hesablama effektivliyi sayəsində böyük verilənlərdə belə tez-tez tətbiq olunur.

In [5]:



1.3. K-Means-in Məqsədi və İstifadə Sahələri 🎯

K-Means-in əsas məqsədi, verilənlər dəstini belə bölməkdir ki, hər klaster daxilində nümunələr bir-birinə mümkün qədər yaxın olsun, digər klasterlərlə isə mümkün qədər uzaq olsun. Bu məqsəd riyazi olaraq **Within-Cluster Sum of Squares (WCSS)** minimallaşdırılması kimi ifadə olunur:

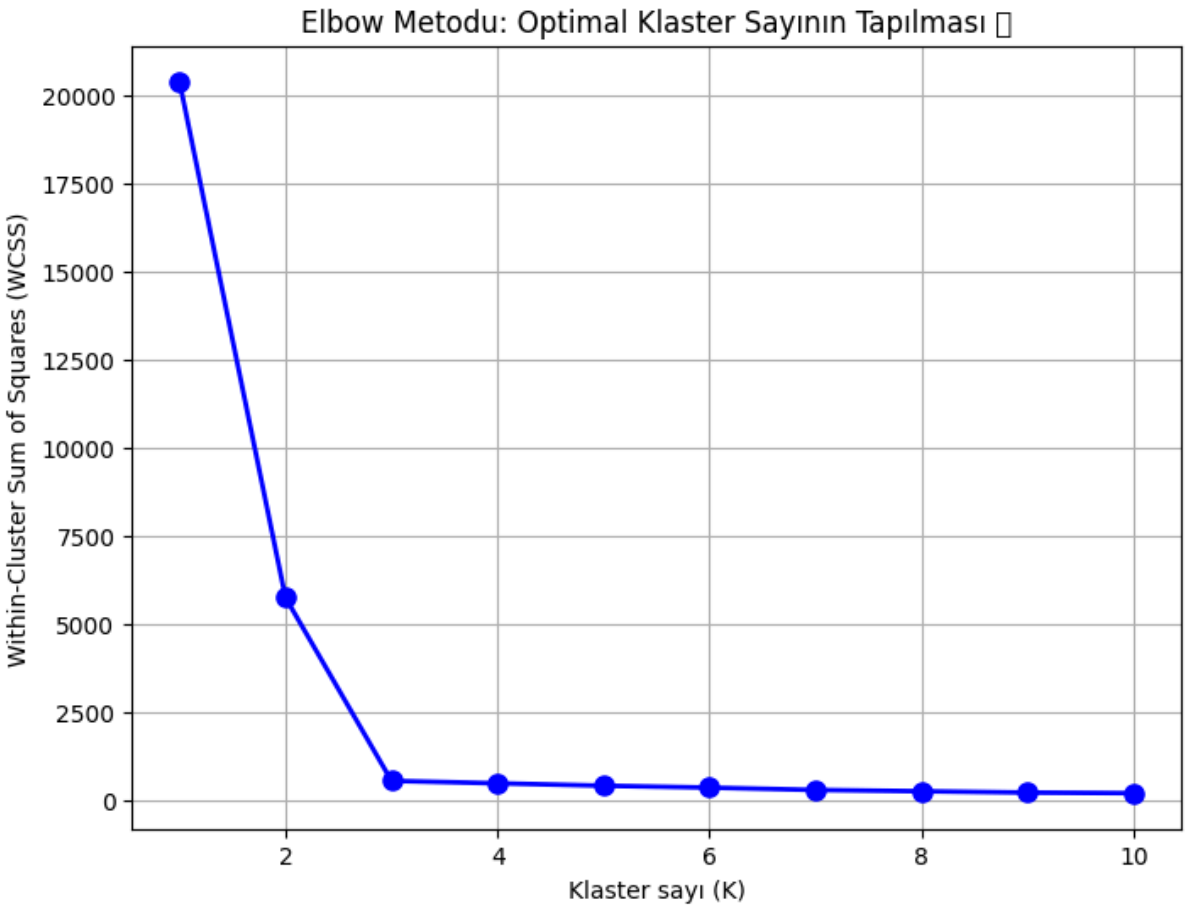
$$\min_{C} \sum_{j=1}^{K} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

Burada C_j — j-ci klaster, x_i klasterə aid olan nümunə və μ_j həmin klasterin centroid-i (orta nöqtəsi) göstərilir.

İstifadə Sahələri:

- **Marketinq və Müştəri Seqmentasiyası:** Müxtəlif müştəri qruplarını taparaq onlara fərqli marketinq strategiyaları tətbiq etmək.
- **Şəkil Emalı:** Rəng kvantizasiyası, yəni böyük rəng palitrasını daha kiçik klasterlərə bölərək yaddaş və sürəti optimallaşdırmaq.
- **Text Mining:** Oxşar mətn sənədlərini qruplaşdırmaq və tematik seqmentasiya aparmaq.
- **Sosial Şəbəkə Analizi:** İstifadəçilərin oxşar davranış və əlaqələrə görə qruplaşdırılması.
- **Biologiya:** Gen ifadə nümunələrinin klasterləşdirilməsi və bioloji funksiyaların araşdırılması.
- **Anomaliya Aşkarlanması:** Normal və anormal nümunələri klasterlərə ayıraraq anomaliyaların tapılması.

In [8]:



1.4. K-Means-in Üstünlükləri və Çatışmazlıqları ✅❌

Üstünlükləri:

- **Sadəlik və Anlaşıqlılıq:** K-Means konsepti intuitivdir və asan tətbiq edilir.
- **Hesablama Effektivliyi:** Xüsusilə böyük verilənlər dəstlərində sürətlə işləyir.
- **Müxtəlif Sahələrdə İstifadə:** Marketinqdən tutmuş bioinformatikaya qədər geniş tətbiq sahəsi var.

Çatışmazlıqları:

- **Klaster Sayının (K) Əvvəlcədən Məlum Olması:** K istifadəçi tərəfindən əvvəlcədən təyin edilməlidir ki, bu da çox vaxt subyektiv və çətin seçim olur.
- **Başlanğıc Mərkəzlərin Həssaslığı:** Təsadüfi başlanğıc nəticənin keyfiyyətinə təsir edir, zəif başlanğıc lokal minimumlara ilişməyə səbəb ola bilər.
- **Klaster Formalarının Məhdudiyyəti:** K-Means sferik (round) və oxşar ölçüdə klasterlər üçün uyğundur, qeyri-sferik və ya müxtəlif ölçülü klasterləri yaxşı modelləşdirə bilmir.
- **Outlierlərə Həssaslıq:** Outlierlər centroid-ləri çəkərək nəticəni poza bilər.
- **Lokal Minimum Problemi:** Alqoritm qlobal optimum deyil, lokal minimum tapmaq ehtimalı yüksəkdir.

2. K-Means Alqoritmasının Riyazi Əsasları 📐

2.1. Verilənlər və Klaster Mərkəzləri

Verilənlər $X = \{x_1, x_2, \ldots, x_n\}$, burada hər x_i d-ölçülü vektordur, yəni $x_i \in \mathbb{R}^d$. Məsələn, müştəri xüsusiyyətləri kimi yaş, gəlir, alış tezliyi kimi d fərqli atribut ola bilər.

Klasterlər $C = \{C_1, C_2, \ldots, C_K\}$ olaraq təsnif edilir, burada hər C_j nümunələr toplusudur. Hər klasterin centroid-i isə klasterə aid nümunələrin vektoral ortalaması kimi hesablanır:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

Bu centroid klasterin "mərkəz nöqtəsi" kimi çıxış edir və növbəti iterasiyada nümunələr ona ən yaxın klaster kimi təyin olunur.

2.2. Ekvlid Məsafəsi və Digər Məsafə Ölçüləri 📏

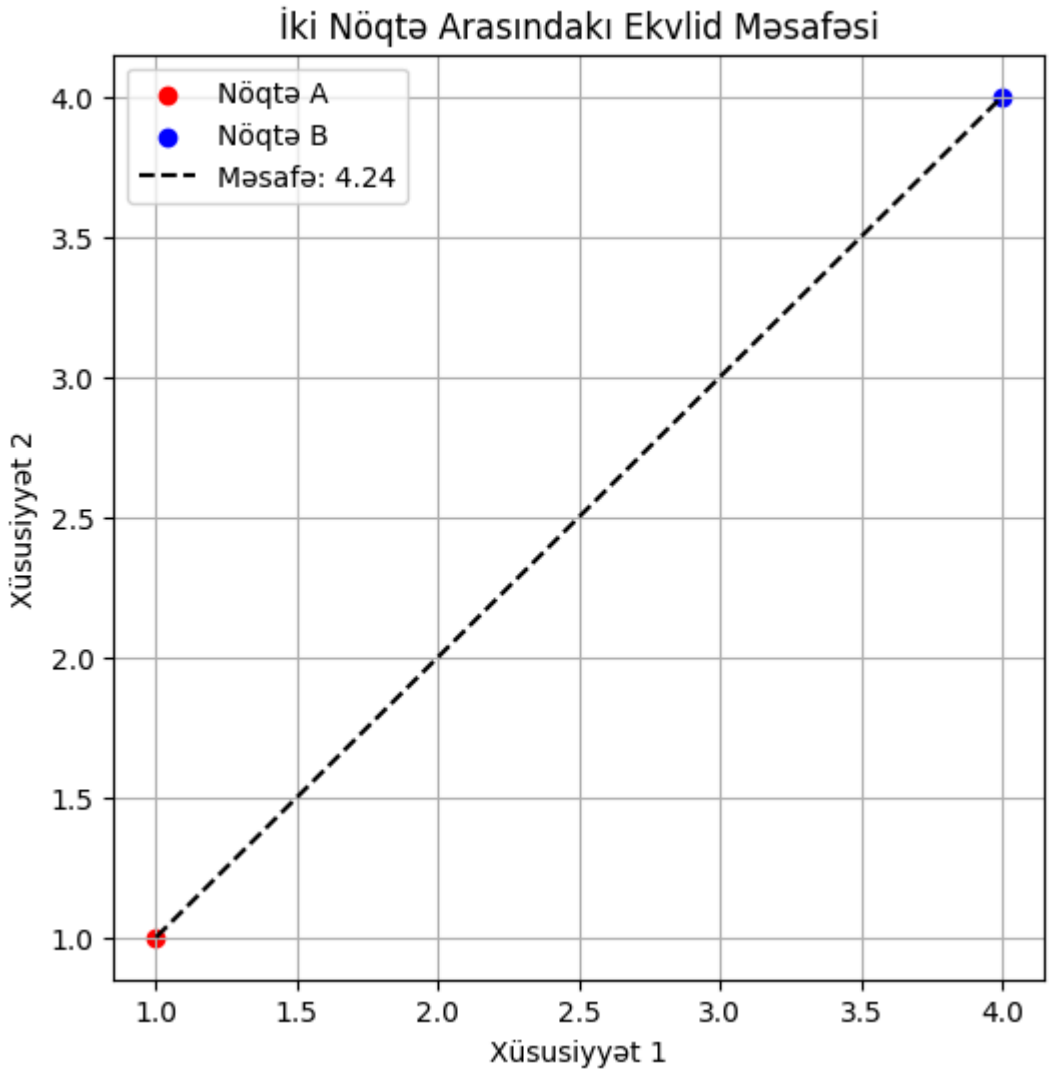
K-Means alqoritmində klasterə aidlığı müəyyən etmək üçün ən çox istifadə olunan məsafə ölçüsü **Euclidean distance**dur:

$$d(x_i, \mu_j) = \sqrt{\sum_{m=1}^{d} (x_{im} - \mu_{jm})^2}$$

Burada x_{im} — i-ci nümunənin m-ci xüsusiyyəti, μ_{jm} isə j-ci klasterin centroid-inin m-ci koordinatıdır. Euclidean distance nöqtələr arasındakı düz xətt məsafəsidir və bu məsafəyə görə ən yaxın centroid müəyyən edilir.

Digər məsafə ölçüləri də mövcuddur (məsələn, Manhattan distance, Cosine similarity), lakin K-Means əsasən Euclidean məsafəsi ilə işləmək üçün nəzərdə tutulub.

In [10]:



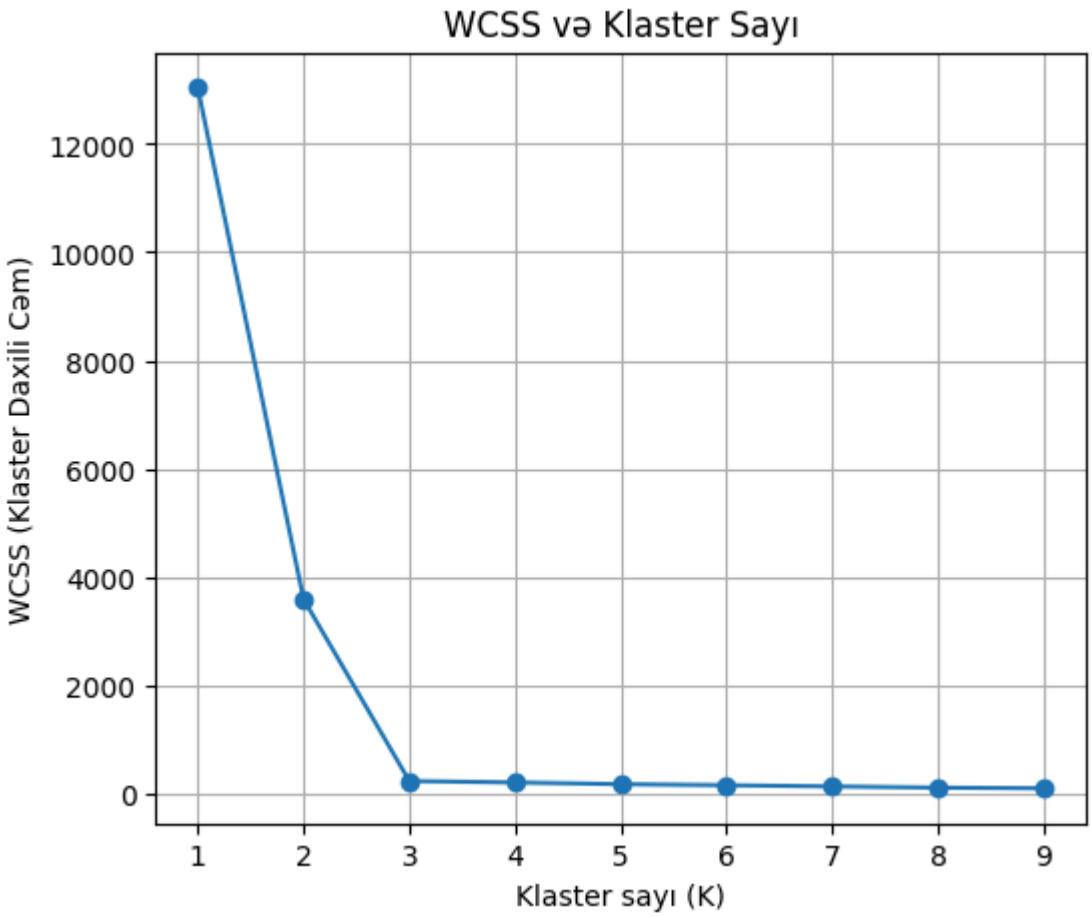
2.3. Optimizasiya Məqsədi (Within-Cluster Sum of Squares)

K-Means alqoritminin əsas məqsədi klaster daxilində nümunələr arasındakı məsafələrin kvadrat cəmini minimallaşdırmaqdır. Riyazi olaraq bu belə ifadə olunur:

$$\min_{C} \sum_{j=1}^{K} \sum_{x_i \in C_j} ||x_i - \mu_j||^2$$

Burada $||x_i - \mu_j||^2$ — x_i və centroid μ_j arasındakı məsafənin kvadratıdır. Bu optimizasiya məqsədi klaster daxilində oxşar nümunələrin sıxlaşmasını təmin edir və klasterlərin arasındakı fərqlərin maksimum olmasını hədəfləyir.

In [11]:



2.4. İterativ Yenilənmə Addımları 🔄

K-Means-in iterativ alqoritmi iki əsas addımdan ibarətdir:

1. **Təyinat Addımı (Assignment Step):**
   Hər nümunə ona ən yaxın centroid-ə aid edilir. Bu mərhələdə məsafələr hesablanır və nümunə üçün ən minimal məsafəyə sahib klaster seçilir:

$$C_j = \{x_i : \|x_i - \mu_j\|^2 \leq \|x_i - \mu_l\|^2, \forall l = 1, \ldots, K\}$$

2. **Yeniləmə Addımı (Update Step):**
   Hər klaster üçün yeni centroid təyin olunur. Bu yeni centroid klasterdəki bütün nümunələrin vektorlarının orta nöqtəsidir:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

Bu iki addım iterativ şəkildə təkrarlanır. Hər iterasiyada nümunələr yenidən klasterlərə aid edilir və centroidlər yenilənir. Proses klaster təyinatları dəyişmədikdə və ya iterasiya limiti keçdikdə dayanır.

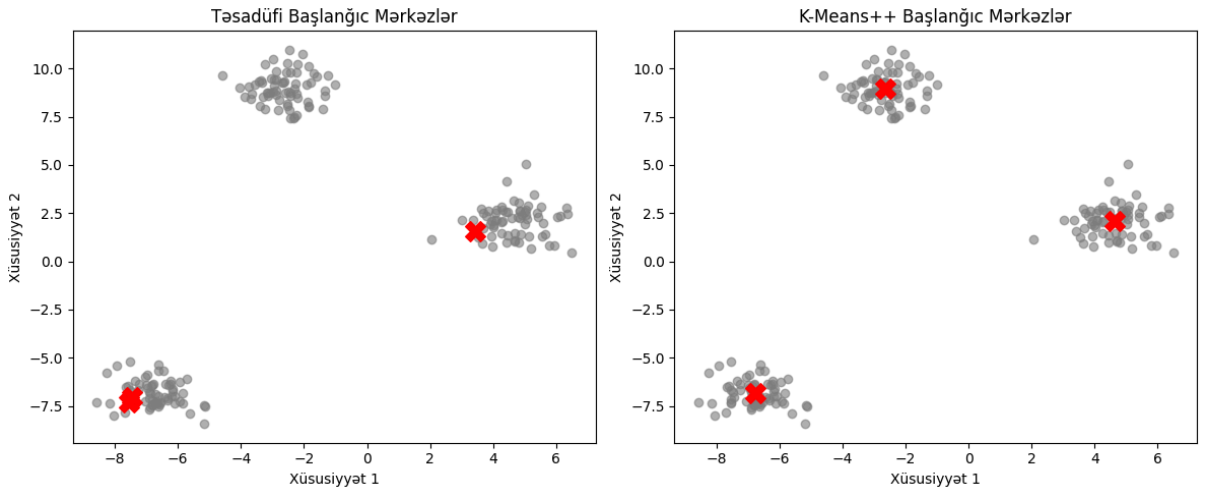
3. K-Means Alqoritmasının İşləmə Prinsipi ⚙️

3.1. Başlanğıc Mərkəzlərin Seçilməsi 🎯

Başlanğıc centroidlərin seçilməsi alqoritmin nəticəsinə güclü təsir göstərir. Əgər başlanğıc nöqtələr pis seçilərsə, alqoritm lokal minimuma ilişə bilər və nəticələr zəif ola bilər. İki əsas yanaşma var:

- **Random Initialization:** Təsadüfi seçilir, lakin bu, zəif nəticələrə səbəb ola bilər.
- **K-Means++ Initialization:** Daha yaxşı başlanğıc nöqtələr seçmək üçün statistik olaraq ən uzaq nöqtələr seçilir. Bu, sürətli konvergensiya və daha stabil nəticələr verir.

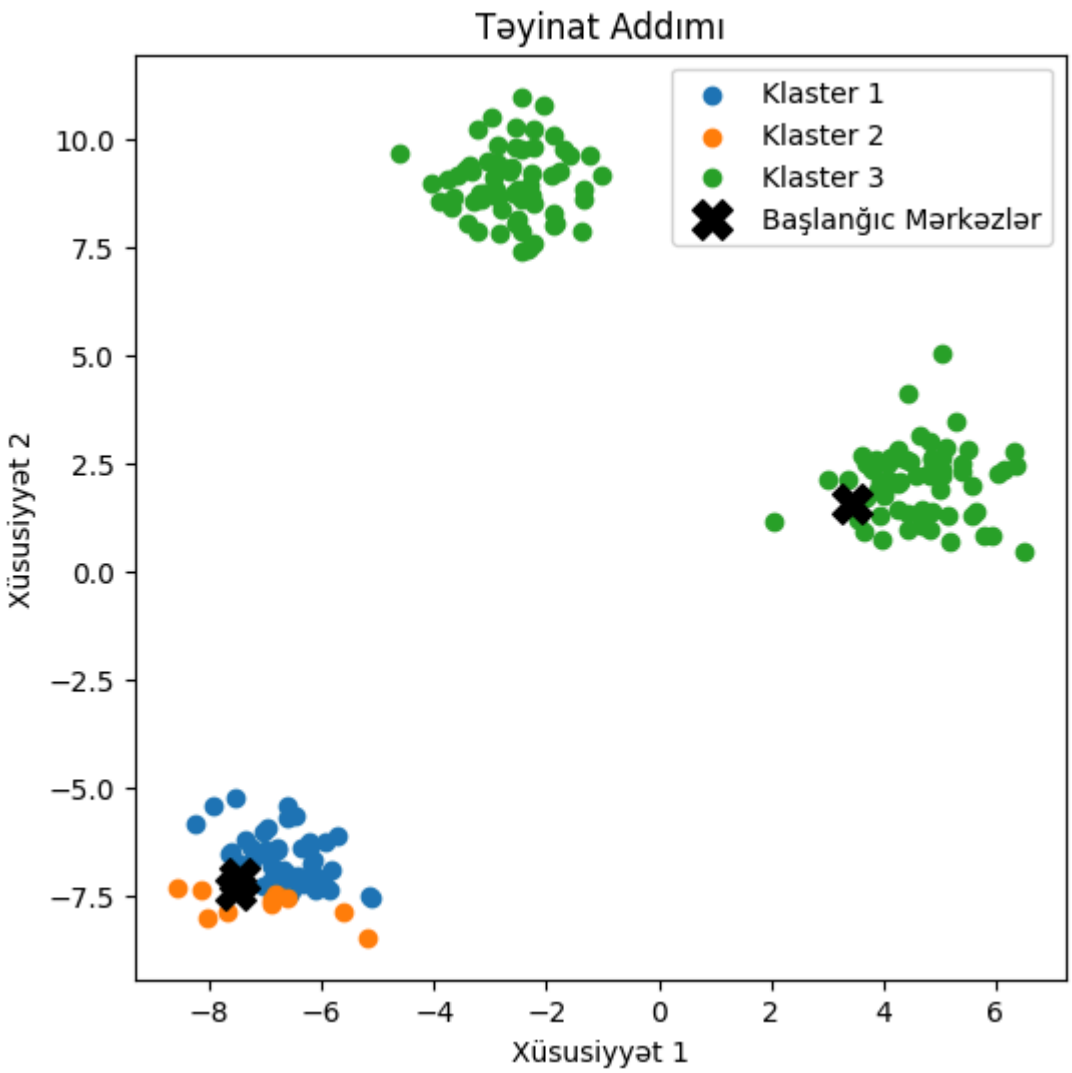
In [18]:



3.2. Təyinat Addımı (Assignment Step)

Bu mərhələdə hər nümunə cari centroidlərdən ən yaxınına aid edilir. Bu addım verilənlərin klasterlərə bölünməsinin əsas mexanizmidir və məsafə ölçüsü çox vaxt Euclidean distance-dır. Nümunələr hər iterasiyada yenidən təyin olunur ki, bu da klasterlərin keyfiyyətini artırır.

In [19]:



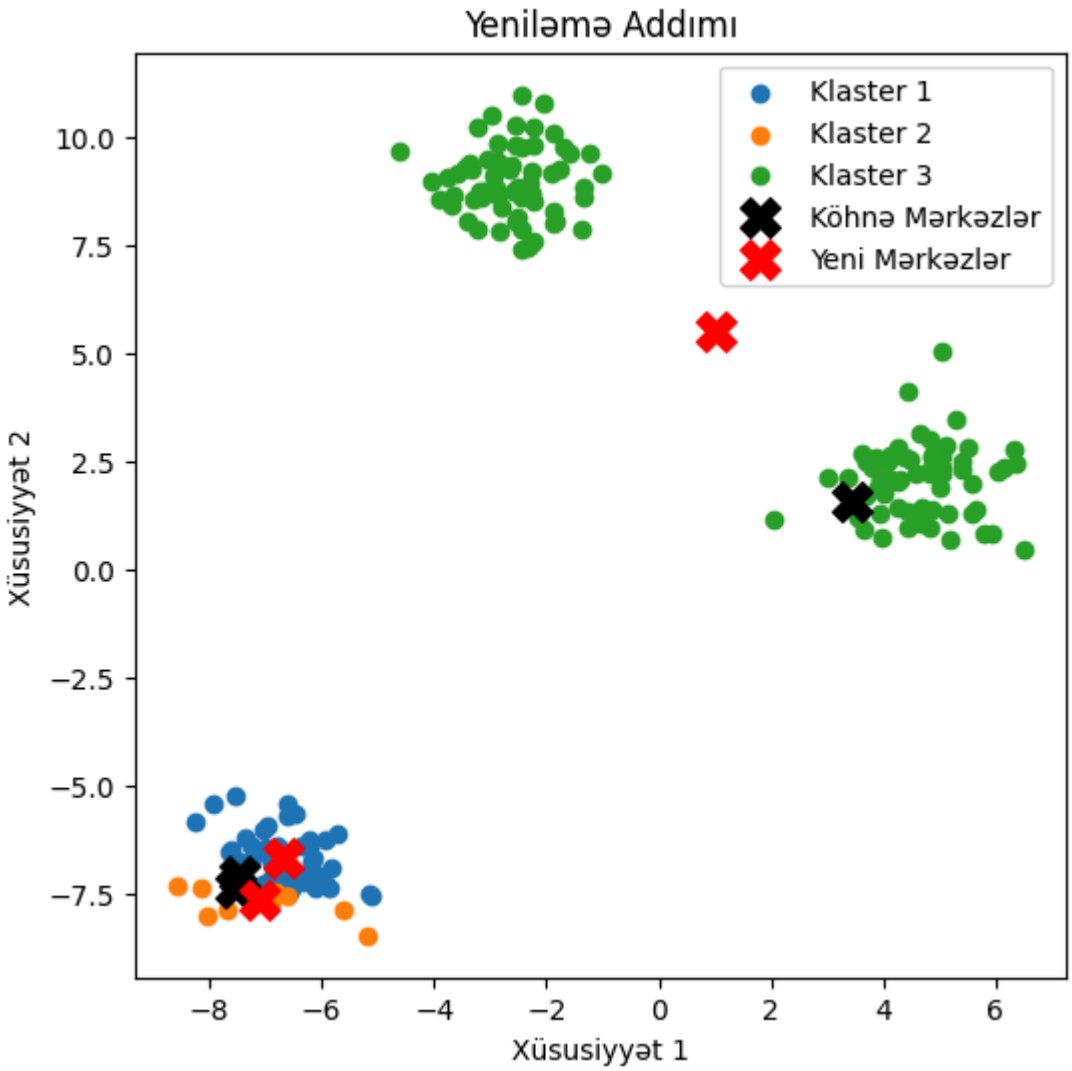
3.3. Yeniləmə Addımı (Update Step) 🔄

Burada hər klaster üçün yeni centroid hesablanır. Yeni centroid klasterdəki bütün nümunələrin vektorlarının orta nöqtəsidir, yəni:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

Bu yeni mərkəz nöqtəsi növbəti iterasiyada təyinat addımında istifadə olunur.

In [20]:



3.4. Dayanma Kriteriyaları və Konvergensiya

Alqoritm aşağıdakı hallarda dayanır:

- Klaster təyinatları əvvəlki iterasiya ilə müqayisədə dəyişmir.
- Centroidlərin dəyişməsi təyin olunmuş kiçik bir toleransdan aşağı düşür.
- Maksimum iterasiya sayı tamamlanır.

Konvergensiya təmin olunmuş olur, yəni nəticədə klaster təyinatları sabitləşir.

4. K-Means Alqoritmasının Tətbiqi 🛠️

4.1. Ümumi İş Axını

K-Means tətbiqinin əsas mərhələləri bunlardır:

1. **Verilənlərin Anlaşılması:**

İlk olaraq verilənlərin strukturu, xüsusiyyətləri və keyfiyyəti dərindən analiz edilir. Bu mərhələdə null dəyərlər, yanlış və ya anormal nöqtələr müəyyənləşdirilir və müvafiq tədbirlər görülür. Məlumatların təbiətini və xüsusiyyətlərini başa düşmək klasterləşdirmənin uğurlu olması üçün vacibdir, çünki keyfiyyətsiz və ya yalnış verilənlər nəticənin etibarlılığını aşağı sala bilər.

2. **Ön İşləmə (Preprocessing):**

Verilənlər miqyaslanır və normallaşdırılır. Çünki K-Means məsafə əsaslı metod olduğundan, xüsusiyyətlərin fərqli ölçüləri klasterlərin formalaşmasına ciddi təsir göstərə bilər. Məsələn, gəlir kimi böyük ölçülü xüsusiyyətlər yaş kimi kiçik ölçülü xüsusiyyətləri "gizlədə" bilər. Buna görə Standartlaşdırma (Z-score), Min-Max scaling və ya digər normallaşdırma üsulları tətbiq olunur. Eyni zamanda itkin dəyərlərin tamamlanması və ya çıxarılması, kənar dəyərlərin (outliers) yoxlanması bu mərhələdə yer alır.

3. **Klaster Sayının Seçilməsi (KKK):**

Ən uyğun klaster sayının təyin edilməsi mühüm addımdır. Bunun üçün Elbow method, Silhouette score kimi analitik və vizual metodlardan istifadə olunur. Doğru seçilmiş K klasterlərin həm mənalı, həm də istifadəyə yararlı olmasını təmin edir.

4. **Başlanğıc Mərkəzlərin Seçilməsi:**

Centroidlərin başlanğıc nöqtələrinin seçilməsi vacibdir. Random seçilmə və ya `K-Means++` kimi daha ağıllı başlanğıc metodları istifadə olunur. Yaxşı başlanğıc, alqoritmin lokal minimuma ilişməsinin qarşısını alır və daha yaxşı nəticələr verir.

5. **Alqoritmin İcrası:**

İterativ şəkildə təyinat (assignment) və yeniləmə (update) addımları icra olunur. Bu addımlar klaster mərkəzləri sabitləşənə və ya maksimum iterasiya sayına çatılana qədər davam edir.

6. **Nəticələrin Qiymətləndirilməsi:**

Tapılan klasterlərin keyfiyyəti müxtəlif metriklərlə ölçülür. Bu mərhələdə daxili qiymətləndirmə (Silhouette score, Davies-Bouldin index) və mümkün olduqda xarici qiymətləndirmə aparılır.

7. **Vizualizasiya:**

Yüksək ölçülü verilənlərdə PCA və ya t-SNE kimi ölçü azaldılması metodları tətbiq edilərək klasterlər iki və ya üç ölçüdə vizuallaşdırılır. Bu, klasterlərin ayrılması və uyğunluğu haqqında intuitiv təsəvvür yaradır.

8. **Praktiki Nəticələrin Interpretasiyası:**

Son mərhələdə tapılan klasterlər iş sahəsinin kontekstində şərh edilir və biznes və ya tədqiqat qərarlarının qəbulunda istifadə olunur. Bu, klasterləşdirmənin real faydasını təmin edir.

4.2. Klaster Keyfiyyəti və Vizualizasiya 🎨

K-Means nəticələrinin nə qədər uğurlu olduğunu qiymətləndirmək üçün müxtəlif metriklərdən istifadə olunur:

- **Silhouette score:**

Hər nümunənin öz klasterinə aidlığı ilə digər ən yaxın klasterə olan məsafələrin nisbətini ölçür. Qiymət -1 ilə $+1$ arasında dəyişir, yüksək dəyər nümunənin düzgün klasterə aid olduğunu göstərir.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

burada $a(i)$ — i-ci nümunənin öz klasterindəki orta məsafəsi, $b(i)$ — i-ci nümunənin digər ən yaxın klasterdəki orta məsafəsidir.

- **Davies-Bouldin index:**

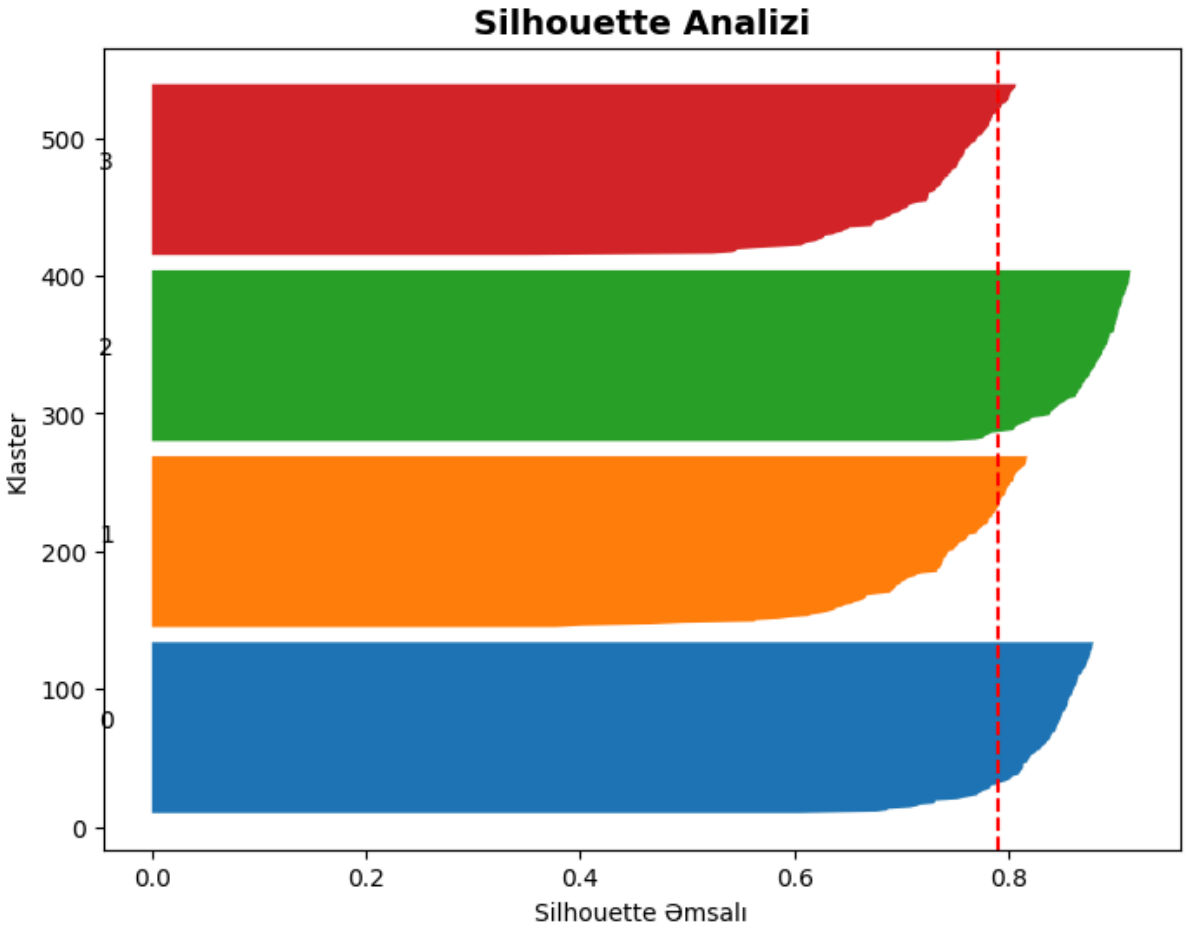
Klasterlərin içərisindəki dispersiya və klasterlərarası məsafələrin nisbətini ölçür. Kiçik dəyər daha yaxşı klasterləşdirməni göstərir.

- **Vizualizasiya:**

Verilənlər çoxölçülü olduqda PCA və ya t-SNE kimi metodlarla ölçü azaldılır və klasterlər 2 və ya 3 ölçüdə vizuallaşdırılır. Bu, klasterlərin real ayrışmasını və uyğunluğunu intuitiv görmək üçün vacibdir.

Bu analizlər nəticələrin güvənilirliyini yoxlamağa və klasterlərin praktiki istifadəsi üçün əlverişliliyini qiymətləndirməyə imkan verir.

In [22]:



5. K-Means Parametrləri və Tənzimləmələri ⚙️

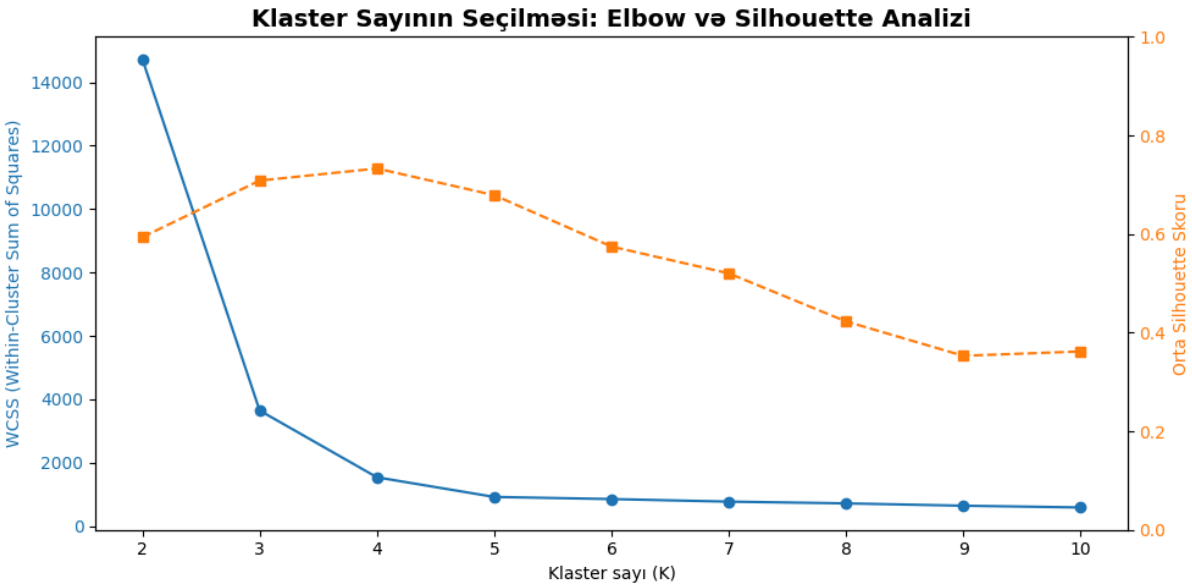
5.1. Klaster Sayının Seçilməsi 🎯

K-Means-də ən kritik parametr K-dır — yəni klaster sayı. Bu, istifadəçinin bilməli olduğu və düzgün təyin etməsi vacib olan parametrdir. Çox kiçik K verilənləri düzgün təmsil etməyə bilər, çox böyük K isə çox sayda kiçik və mənasız klasterə səbəb ola bilər.

Klaster sayını seçmək üçün əsas metodlar:

- **Elbow method:** K dəyişdikcə WCSS qiyməti azalmağa davam edir, lakin bir nöqtədə azalma tempində kəskin yavaşlama olur (dirsək nöqtəsi). Bu nöqtə optimal K kimi qəbul edilir.
- **Silhouette analysis:** Müxtəlif K üçün Silhouette skorları hesablanır və ən yüksək orta skor verən K seçilir.
- **Davies-Bouldin index** və digər metriklər də istifadə edilə bilər.

In [24]:



5.2. Başlanğıc Mərkəzlərin Seçilməsi və Önəmi 🔀

Başlanğıc mərkəzlərin seçilməsi nəticənin stabilliyini və keyfiyyətini birbaşa təsir edir. Təsadüfi başlanğıc bəzən zəif nəticələrə səbəb olur. Buna görə `K-Means++` metodu təklif olunur. Burada başlanğıc centroid-lər daha "ağıllı" seçilir, yəni ilk nöqtə təsadüfi götürülür, sonrakılar isə əvvəlkilərdən mümkün qədər uzaq nöqtələr arasından seçilir. Bu metod sürətli konvergensiya və daha yaxşı nəticələr verir.

5.3. Maksimum İterasiya Sayı və Tolerans ⏳

Alqoritmin dayandırılması üçün maksimum iterasiya sayı təyin olunur (məsələn, 300 iterasiya). Həmçinin centroidlərin dəyişməsi çox kiçik olduqda (tolerans səviyyəsi) alqoritm dayanır. Bu, hesablama resurslarının boş yerə sərf olunmasının qarşısını alır.

5.4. İşin Təkrarlanması (n_init) 🔁

Çünki başlanğıc nöqtələr təsadüfi seçildiyindən alqoritm fərqli işlərdə fərqli nəticələr verə bilər. Buna görə K-Means `n_init` parametrinə malikdir ki, burada alqoritm bir neçə dəfə başlanğıc nöqtələrini dəyişərək işə salınır və ən yaxşı nəticə seçilir. Bu təkrarlama prosesi nəticələrin sabitliyi və keyfiyyətini artırır.

6. Klaster Sayının Müəyyənləşdirilməsi Metodları 🧮

Düzgün klaster sayını seçmək çətin məsələdir və bu, klasterləşdirmə performansını ciddi şəkildə təsir edir. Aşağıdakı metodlar ən çox istifadə olunur:

- **Elbow Method:**
  Daxili variasiyanın (WCSS) K-yə görə dəyişməsini göstərir. K artdıqca WCSS azalır, lakin bir nöqtədən sonra azalmanın tempi kəskin azalır. Bu nöqtə dirsək (elbow) nöqtəsi sayılır.

$$WCSS = \sum_{j=1}^{K} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

- **Silhouette Score:**
  Hər nümunənin öz klasteri ilə digər ən yaxın klaster arasında məsafə fərqini ölçür. Orta Silhouette score yüksək olduqda klasterlər yaxşı ayrılmış olur.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

burada $a(i)$ — i-ci nümunənin öz klasterindəki orta məsafəsi, $b(i)$ — i-ci nümunənin digər ən yaxın klasterdəki orta məsafəsidir.

- **Davies-Bouldin Index:**
  Klasterlərin içərisindəki dispersiyanın və klasterlərarası məsafələrin nisbətini qiymətləndirir. Kiçik dəyər daha yaxşıdır.

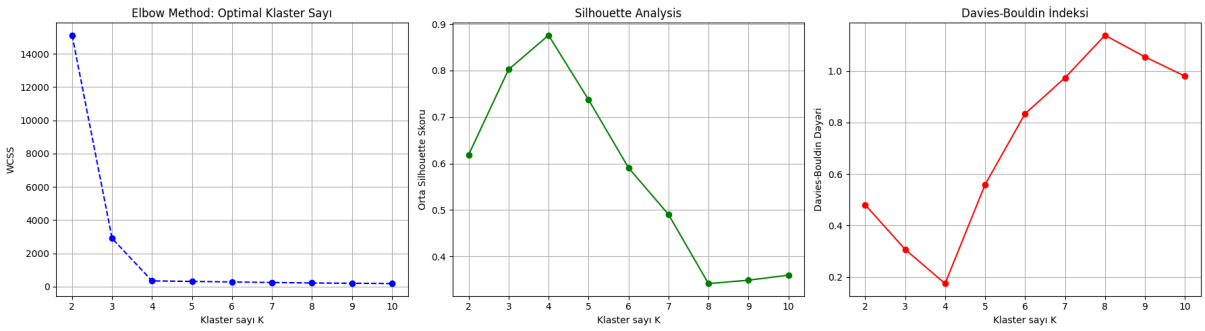
$$DB = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right)$$

burada S_i — klaster i-nin dispersiyası, M_{ij} — klasterlərarası məsafə.

- **Gap Statistic:**
  Verilənlərin təsadüfi paylanması ilə müqayisə apararaq optimal K-ni müəyyən edir.

- **Domen Bilgiləri:**
  Bəzən iş sahəsinə əsaslanaraq K dəyəri təyin olunur.

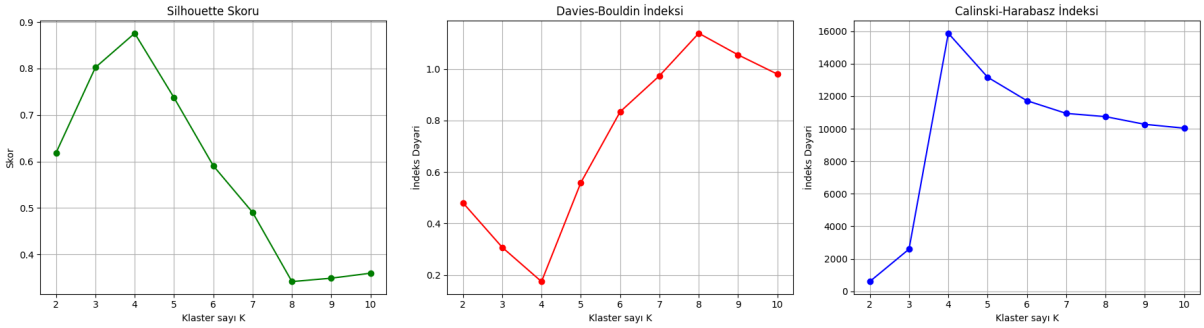
In [29]:



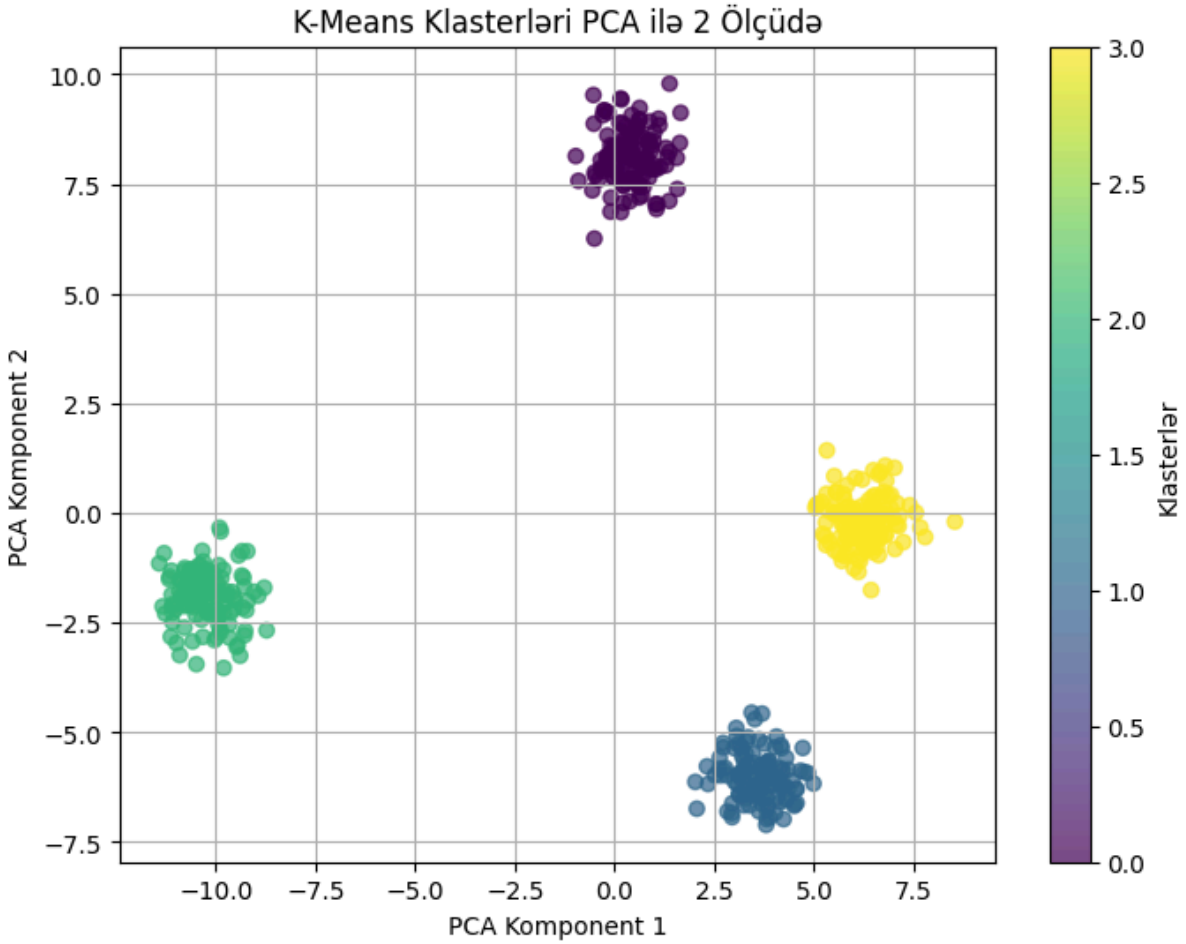
7. K-Means Performansının və Nəticələrinin Qiymətləndirilməsi 📊

Klasterlərin keyfiyyəti həm daxili metriklərlə, həm də xarici məlumat varsa, uyğunluq qiymətləndirmələri ilə ölçülə bilər. Silhouette score, Davies-Bouldin index, Calinski-Harabasz indeksi kimi göstəricilərdən istifadə olunur. Məlumat çox ölçülü olduqda, vizualizasiya üçün PCA və t-SNE kimi metodlarla klasterlərin ayrılması baxılır.

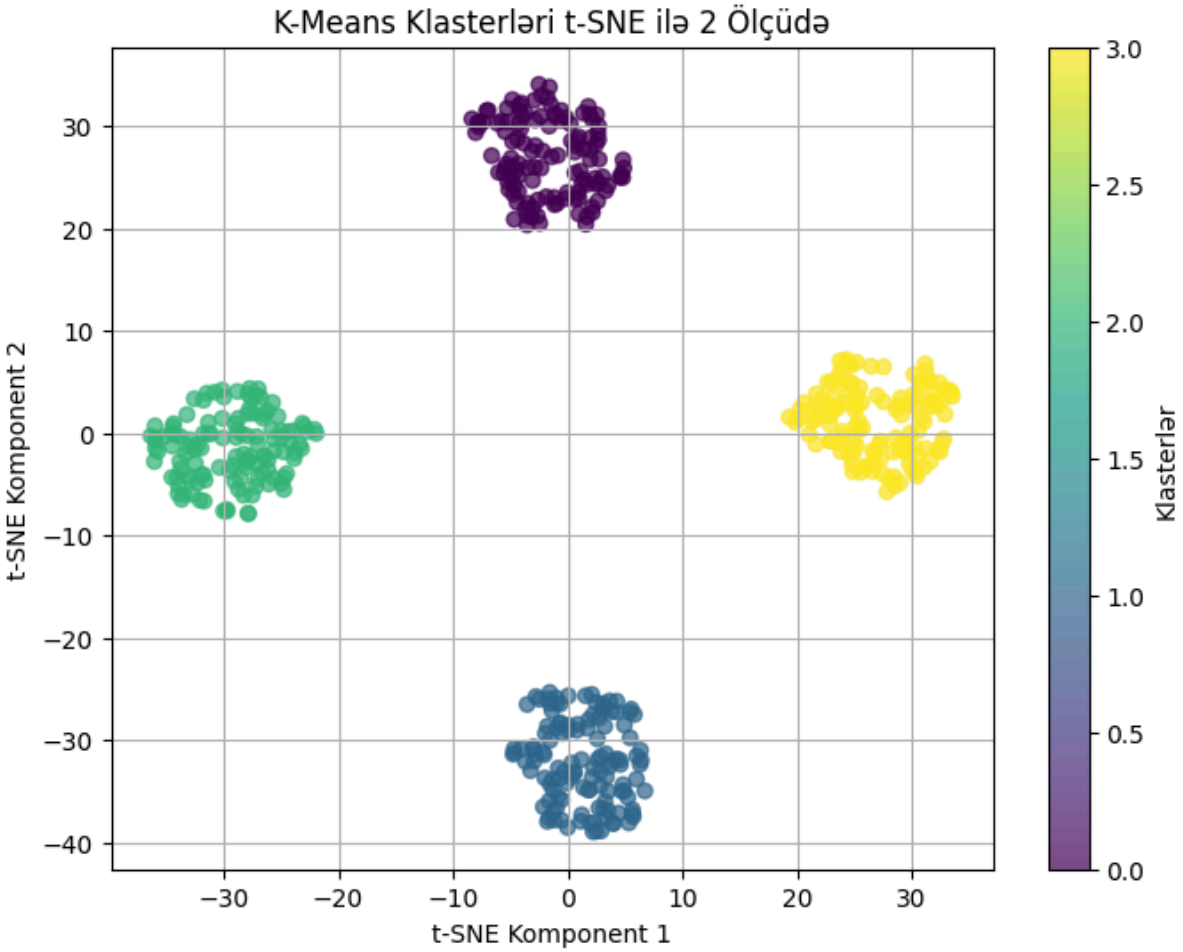
In [30]:


Silhouette Skoru
Skor
Klaster sayı K
Davies-Bouldin İndeksi
İndeks Dəyəri
Klaster sayı K
Calinski-Harabasz İndeksi
İndeks Dəyəri
Klaster sayı K

In [31]:


K-Means Klasterləri PCA ilə 2 Ölçüdə
PCA Komponent 2
PCA Komponent 1
Klasterlər

In [32]:



8. K-Means-in Məhdudiyyətləri və Problemləri ⚠️

K-Means-in bəzi əsas çatışmazlıqları var:

- **Sferik Klasterlər Üçün Uyğundur:** K-Means klasterlərin forması kimi sferik, oxşar ölçülü olmasını tələb edir. Düzbucaqlı, sıx olmayan və ya dəyişkən formalı klasterlərdə zəif nəticə verir.
- **Outlierlərə Həssasdır:** Outlierlər centroid-ləri çəkərək klasterlərin formasını və yerləşməsini poza bilər.
- **Lokal Minimum Problemi:** Alqoritm qlobal optimal həll yerinə, başlanğıc nöqtələrdən asılı olaraq lokal minimuma ilişə bilər.
- **Xətti Olmayan Sərhədləri Modelləşdirməkdə Zəifdir:** K-Means sərhədləri xətti və ya konveksdir, qeyri-xətti sərhədləri ayırmaq üçün yararsızdır.

9. K-Means Alternativləri və Təkmilləşdirmələr 🔄

K-Means-in məhdudiyyətlərini aradan qaldırmaq üçün bir sıra alternativ və təkmilləşdirilmiş metodlar mövcuddur:

- **K-Medoids:**
  Centroidlər əvəzinə, klasterlərin mərkəzində verilənlərdən biri (medoid) seçilir. Bu, outlier dəyərlərə daha davamlıdır.

- **Gaussian Mixture Models (GMM):**
  Verilənlərin müxtəlif Gaussian paylanmalarının qarışığı kimi modelləşdirir və qeyri-sferik klasterləri daha yaxşı əhatə edir.

- **DBSCAN (Density-Based Spatial Clustering):**
  Sıxlıq əsaslı klasterləşdirmədir, klasterlərin formasını əvvəlcədən təyin etmir və anomaliyaları aşkarlaya bilir.

- **Spectral Clustering:**
  Qraf nəzəriyyəsindən istifadə edərək mürəkkəb və qeyri-xətti sərhədləri model edir.

- **Mini-Batch K-Means:**
  Böyük verilənlər üçün daha sürətli işləyən K-Means variantıdır, kiçik təsadüfi alt nümunələrlə iterasiya edir.

Bu metodlar K-Means-in zəif tərəflərini azaltmaq və müxtəlif tətbiq sahələrində daha yaxşı nəticələr əldə etmək üçün istifadə olunur.

10.K-Means-in Real Dünya Tətbiqləri 🌍

K-Means geniş real dünya tətbiqlərinə malikdir:

- **Müştəri Seqmentasiyası:**
  Müxtəlif davranış və ehtiyaclara malik müştəri qruplarının müəyyənləşdirilməsi. Bu, marketinq strategiyalarını fərdiləşdirmək və müştəri məmnuniyyətini artırmaq üçün istifadə olunur.

- **Şəkil Emalı:**
  Rənglərin və ya şəkil bölgələrinin qruplaşdırılması. Məsələn, rəng kvantizasiyası və ya şəkil seqmentasiyasında K-Means effektivdir.

- **Anomaliya Aşkarlanması:**
  Normal və anormal nümunələrin ayrılması üçün ilkin mərhələ kimi istifadə olunur. Anormal nümunələr klasterlərdən kənarda qaldığı üçün aşkarlanır.

- **Text Mining:**
  Oxşar mətn sənədlərinin qruplaşdırılması və analiz edilməsi, məsələn, mövzu əsaslı sənəd qruplarının yaradılması.

- **Sosial Şəbəkə Analizi:**
  İstifadəçi davranışlarının qruplaşdırılması və sosial qrupların aşkarlanması üçün istifadə olunur. Bu, sosial əlaqələrin və təsirlərin öyrənilməsində faydalıdır.

11. K-Means ilə Bağlı və Tövsiyələr 💡

11.1. Data Preprocessing: standartazing və Scaling ⚖️

Özəlliklərin miqyası fərqli olduqda, böyük miqyası olan xüsusiyyətlər klaster təyinatına daha çox təsir edə bilər. Məsələn, gəlir 1000-lərlə, yaş isə onluqlarla ifadə olunursa, gəlir klaster təyinatını dominant edəcək. Buna görə verilənlər standartlaşdırılır (məsələn, Z-score standartlaşdırması):

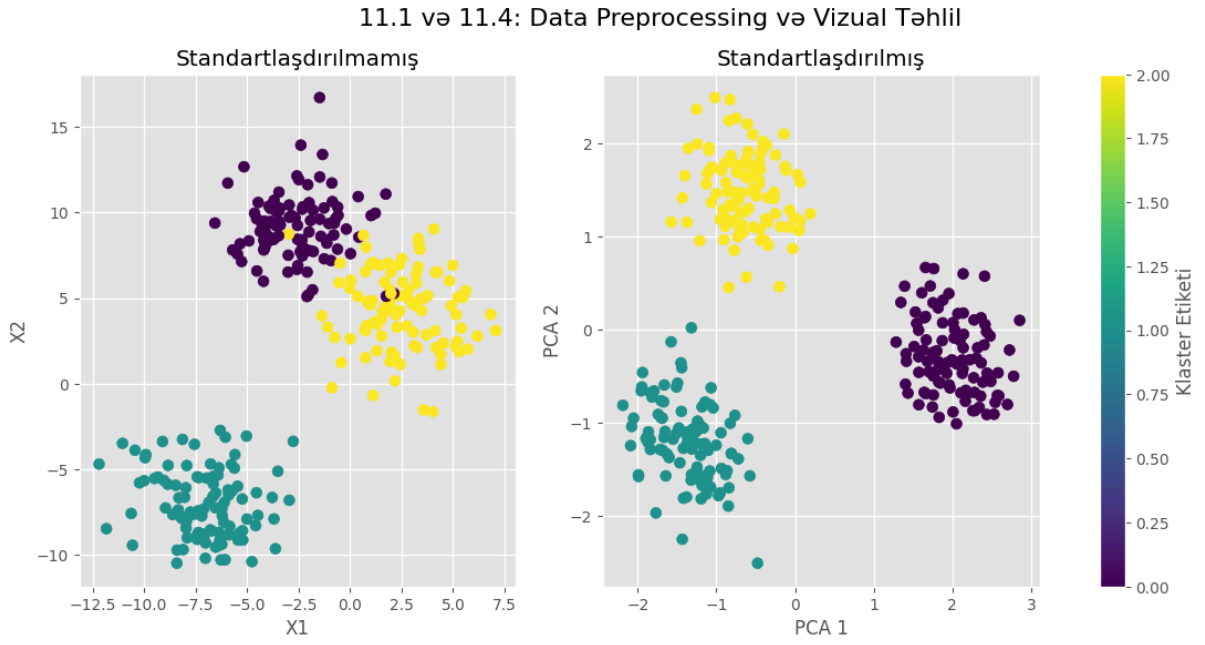
$$z = \frac{x - \mu}{\sigma}$$

Bu yolla bütün xüsusiyyətlər oxşar ölçüyə gətirilir və K-Means-in məsafə hesabları balanslaşdırılır.

11.2. Feature Selection və Engineering 🛠️

Klasterləşdirmənin keyfiyyəti çox vaxt istifadə olunan xüsusiyyətlərdən asılıdır. Informativ, klasterləri yaxşı ayıran xüsusiyyətlərin seçilməsi vacibdir. Bunun üçün statistik analiz, korrelyasiya yoxlanışı və ya dimensionality reduction metodlarından istifadə edilə bilər.

In [47]:



11.3. Çoxsaylı Cəhd ilə Stabilizasiya 🔁

Təsadüfi başlanğıc nöqtələrinin nəticəyə təsirini azaltmaq üçün K-Means bir neçə dəfə fərqli başlanğıc nöqtələri ilə işə salınır və ən yaxşı nəticə seçilir. Bu təcrübə alqoritmin daha stabil və keyfiyyətli nəticələr verməsinə səbəb olur.

11.4. Nəticələrin Vizual Təhlili və Interpretasiyası 🖼️

Tapılan klasterlər iki və ya üç ölçüyə endirilərək PCA və ya t-SNE kimi metodlarla vizuallaşdırılır. Bu, klasterlərin real ayrışmasını görməyə, onların uyğunluğunu qiymətləndirməyə və mümkün problemləri aşkar etməyə imkan verir. Əldə olunan nəticələr iş sahəsinin kontekstində şərh olunmalı və qərarların qəbulunda istifadə edilməlidir.