Makine Öğrenimi ile Elektrik Tüketim Tahmini

Veri Analizi ve Modelleme

Sinem Şen
Beşir Kebbe
Özgür Güvercin

Proje Tanımı

Projenin Amacı: Bu proje, bir şehirdeki 1 yıllık toplam elektrik tüketimini tahmin etmeyi amaçlamaktadır.

Projenin Kapsamı: Bu proje, veri analitiği ve makine öğrenimi tekniklerini kullanarak elektrik tüketimi tahmini yapmayı içermektedir.

Proje Hedefleri:

- Toplam elektrik tüketimini tahmin etmek için doğru ve güvenilir bir model geliştirmek.
- Tahmin modeli için en uygun algoritmaları ve özellikleri belirlemek.
- Gelecek yılın elektrik talebini etkileyen faktörleri belirlemek ve analiz etmek.
- Doğru tahminler yaparak enerji planlaması ve kaynak yönetimi için verimli stratejiler geliştirmek.

Proje Tanımı

Kullanılan Veriler:

- Elektrik tüketimi verileri (aylık/total)
- Bina Özellikleri (Bina Tipi, Bina Alt Tipi, Ortalama Kat Sayısı, Ortalama Bina Yaşı, Ortalama Ev Büyüklüğü)
- Nüfus Verileri (Toplam Nüfus)
- Coğrafi Konum Verileri (İlçeler)

Beklenen Sonuçlar:

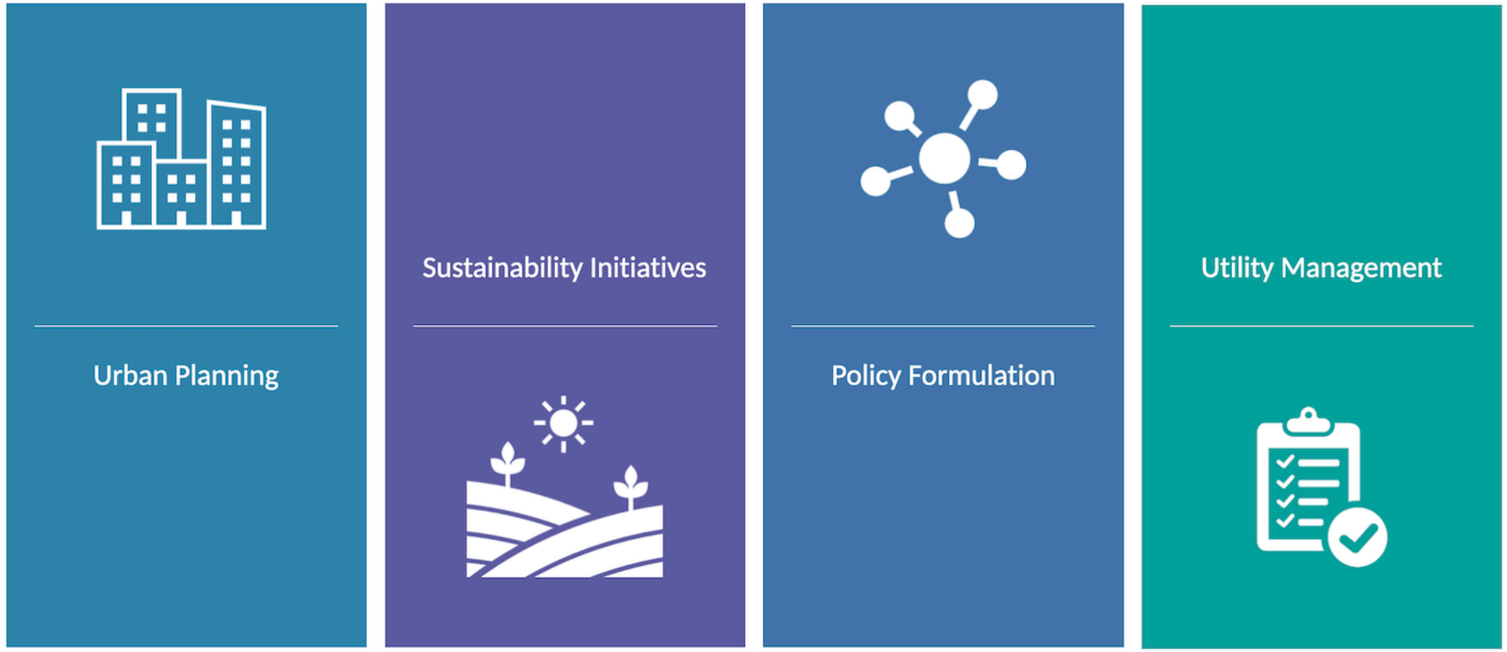
- Gelecek yılın toplam elektrik tüketimini doğru bir şekilde tahmin eden bir model geliştirmek.
- Hangi faktörlerin elektrik tüketimini etkilediğini ve ne kadar etkilediğini belirlemek.
- Doğru tahminler yaparak enerji planlaması ve altyapı yatırımları için verimli stratejiler geliştirmek.

Tahmin Sonucumuzun Önemi

Enerji Planlaması: Bir şehir, bölge veya ülkenin gelecekteki enerji talebini doğru bir şekilde tahmin etmek, elektrik arzını ve dağıtımını planlamak için gereklidir.

Altyapı Yatırımları: Yatırım yapılacak bölgelerin belirlenmesi, yeni santrallerin veya dağıtım hatlarının inşa edilmesi ve mevcut altyapının güçlendirilmesi gibi kararlar, tahminlere dayanarak yapılır.

Fiyatlandırma Stratejileri: Elektrik tüketimi tahminleri, enerji fiyatlarının belirlenmesinde ve talep tabanlı fiyatlandırma stratejilerinin geliştirilmesinde kullanılır. Talebin yoğun olduğu saatlerde veya dönemlerde fiyatların artırılması, talebin dengeye kavuşturulmasına yardımcı olabilir.



Tahmin Sonucumuzun Önemi

Yenilenebilir Enerji Entegrasyonu: Rüzgar türbinlerinin veya güneş panellerinin elektrik üretimini, tahmini talep düzeylerine göre planlamak mümkündür.

Enerji Verimliliği: Belirli bir bölgede elektrik tüketiminin artacağı tahmin ediliyorsa, enerji tasarrufu sağlayacak projelerin başlatılması planlanabilir.

Dataset Genel Bakış

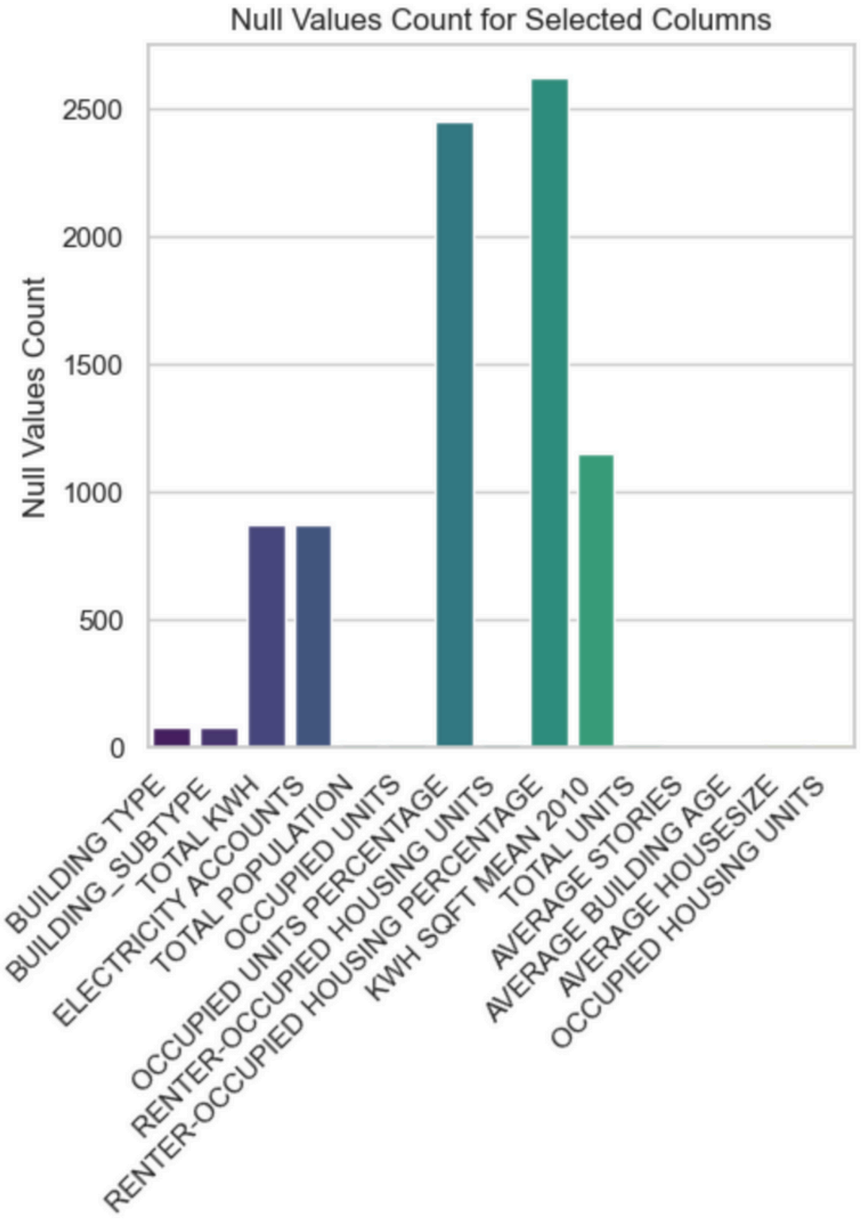
Veri kümesi, Chicago'da 2010 yılı için toplam elektrik ve gaz tüketimi, konut birimi doluluğu, demografik bilgiler, bina özellikleri ve daha fazlasını içeren çeşitli metrikleri kapsar.

Veri kümesi 67,051 satır içeriyor ve 73 özellik kapsıyor.

Veri kümesinin toplam boyutu: 24.7MB

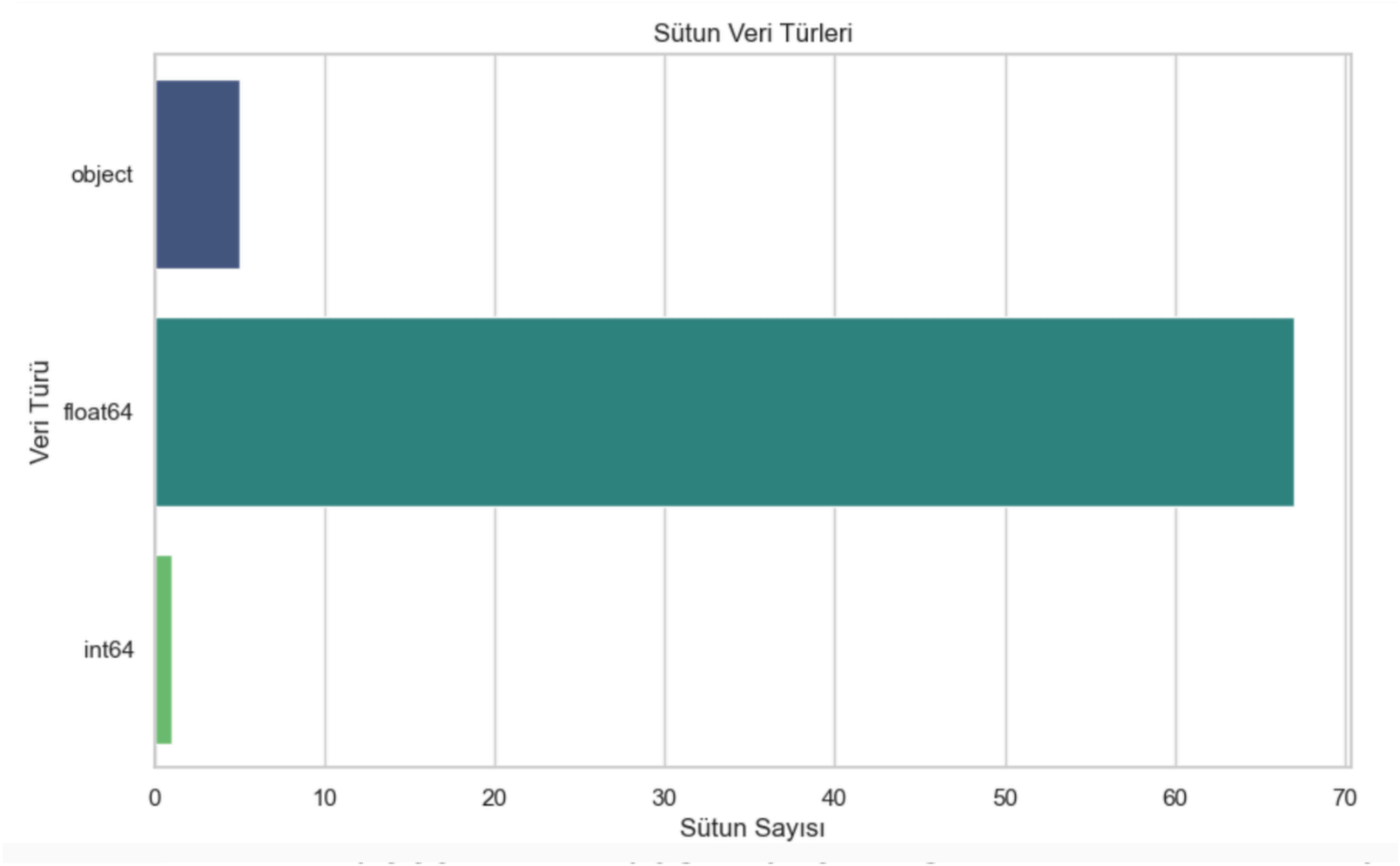
[https://data.cityofchicago.org/Environment-Sustainable-Development/Energy-Usage-2010/8yq3-m6wp/data]
(https://data.cityofchicago.org/Environment-Sustainable-Development/Energy-Usage-2010/8yq3-m6wp/data)

Veri Analizi ve Temizleme: Boş Değerlerin Tespiti ve Temizlenmesi




```
COMMUNITY AREA NAME                  0
CENSUS BLOCK                         0
BUILDING TYPE                        0
BUILDING_SUBTYPE                     0
KWH JANUARY 2010                     0
KWH FEBRUARY 2010                    0
KWH MARCH 2010                       0
KWH APRIL 2010                       0
KWH MAY 2010                         0
KWH JUNE 2010                        0
KWH JULY 2010                        0
KWH AUGUST 2010                      0
KWH SEPTEMBER 2010                   0
KWH OCTOBER 2010                     0
KWH NOVEMBER 2010                    0
KWH DECEMBER 2010                    0
TOTAL KWH                            0
ELECTRICITY ACCOUNTS                 0
ZERO KWH ACCOUNTS                    0
KWH MEAN 2010                        0
KWH MINIMUM 2010                     0
KWH 1ST QUARTILE 2010                0
KWH 2ND QUARTILE 2010                0
KWH 3RD QUARTILE 2010                0
KWH MAXIMUM 2010                     0
KWH SQFT MEAN 2010                   0
TOTAL POPULATION                     0
TOTAL UNITS                          0
AVERAGE STORIES                      0
AVERAGE BUILDING AGE                 0
AVERAGE HOUSESIZE                    0
OCCUPIED UNITS                       0
OCCUPIED UNITS PERCENTAGE            0
RENTER-OCCUPIED HOUSING UNITS        0
RENTER-OCCUPIED HOUSING PERCENTAGE   0
OCCUPIED HOUSING UNITS               0
dtype: int64
```

Veri Analizi ve Temizleme: Kategorik Verilerin Tespiti

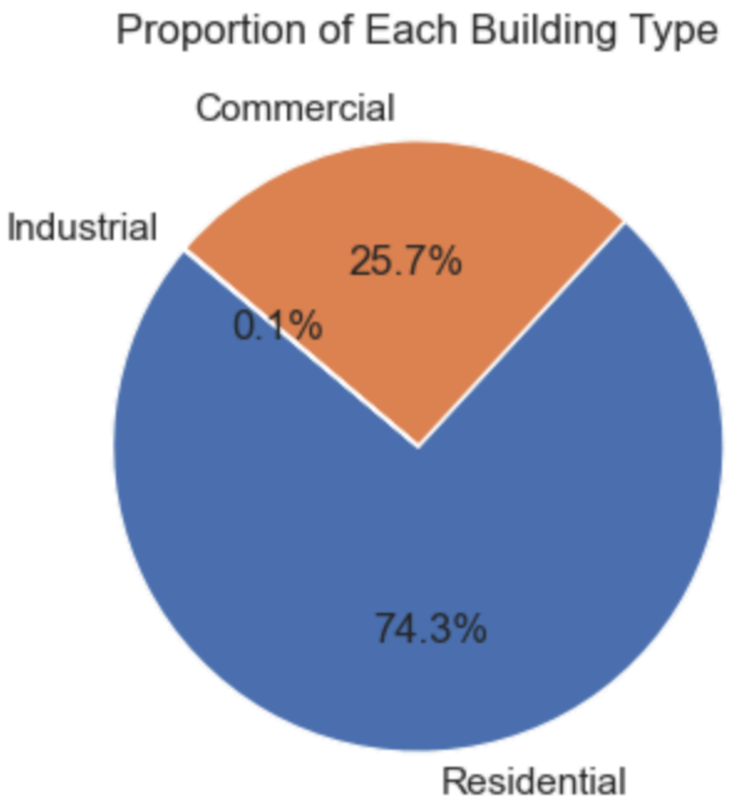



```
dtypes: float64(67), int64(1), object(5)
memory usage: 37.3+ MB
```

Veri Analizi ve Temizleme: Kategorik Verilerin Düzeltilmesi

```
CENSUS BLOCK                          float64
KWH JANUARY 2010                      float64
KWH FEBRUARY 2010                     float64
KWH MARCH 2010                        float64
KWH APRIL 2010                        float64
                                       ...
ELECTRICITY ACCOUNTS_92                 bool
ELECTRICITY ACCOUNTS_94                 bool
ELECTRICITY ACCOUNTS_96                 bool
ELECTRICITY ACCOUNTS_97                 bool
ELECTRICITY ACCOUNTS_Less than 4        bool
Length: 232, dtype: object
Kategorik veri içeren sütun bulunmuyor.
```

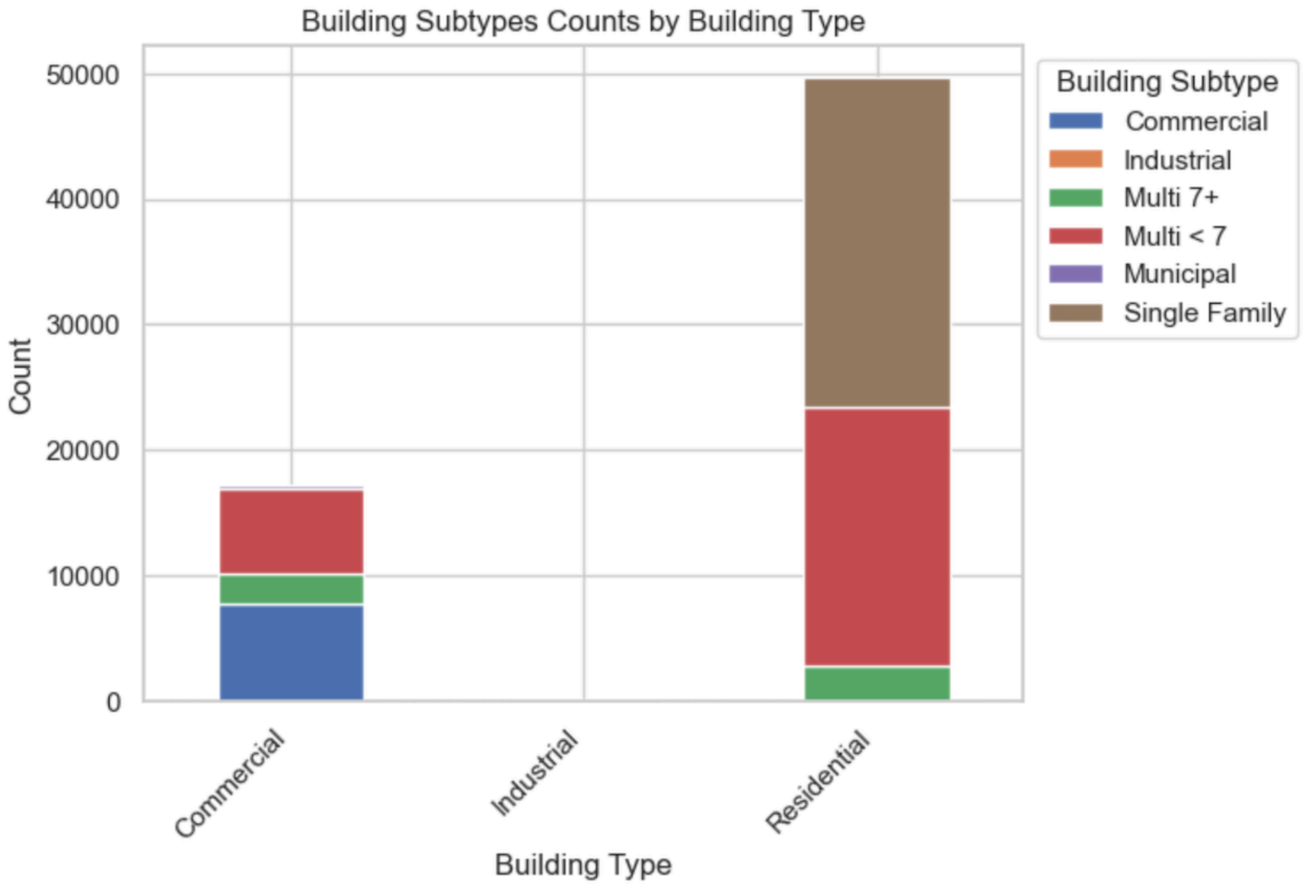
Veri Analizi ve Temizleme: Veri İnceleme



Veri setimizde üç bina türü bulunmaktadır: Konut, Ticari, Endüstriyel.

Toplam 67,051 kayıttan, 49,747'si konut, 17,185'i ticari ve geri kalan 47 kayıt endüstriyel özelliklere aittir.

Veri Analizi ve Temizleme: Veri İnceleme



Ticari:

- Ticari: 7775
- Çoklu < 7: 6731
- Çoklu 7+: 2396
- Belediye: 282
- Tek Aile: 1

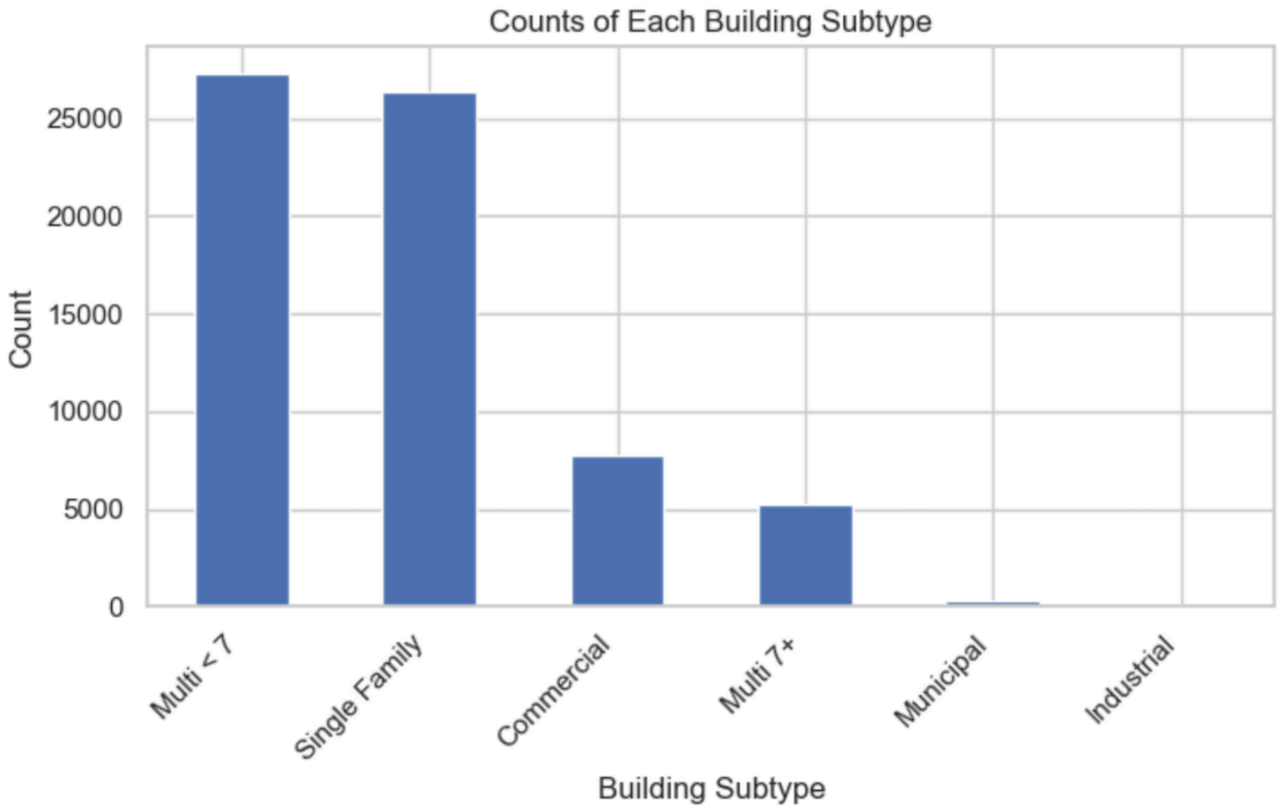
Endüstriyel:

- Endüstriyel: 42

Konut:

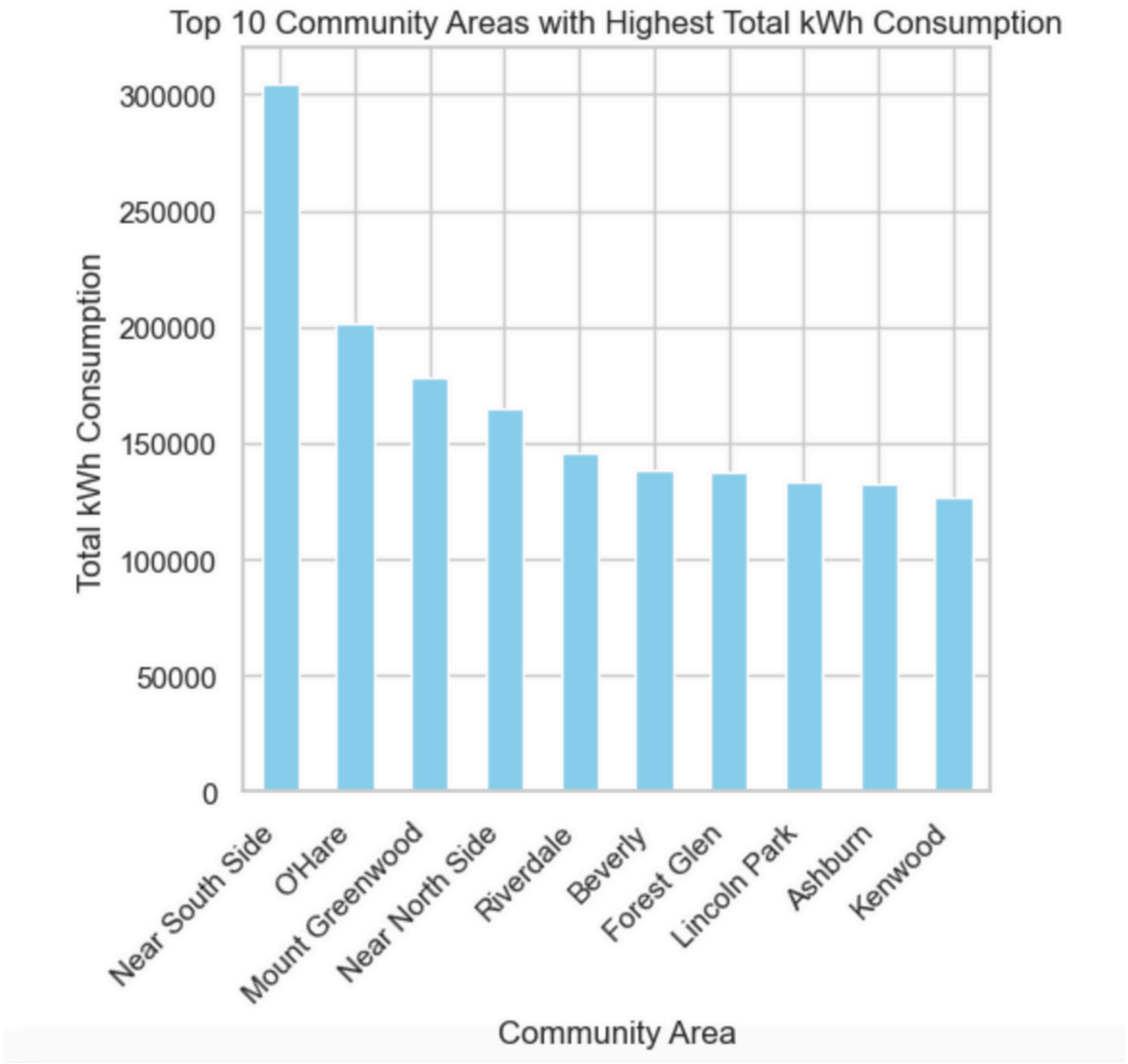
- Tek Aile: 26365
- Çoklu < 7: 20553
- Çoklu 7+: 2829

Veri Analizi ve Temizleme: Veri İnceleme

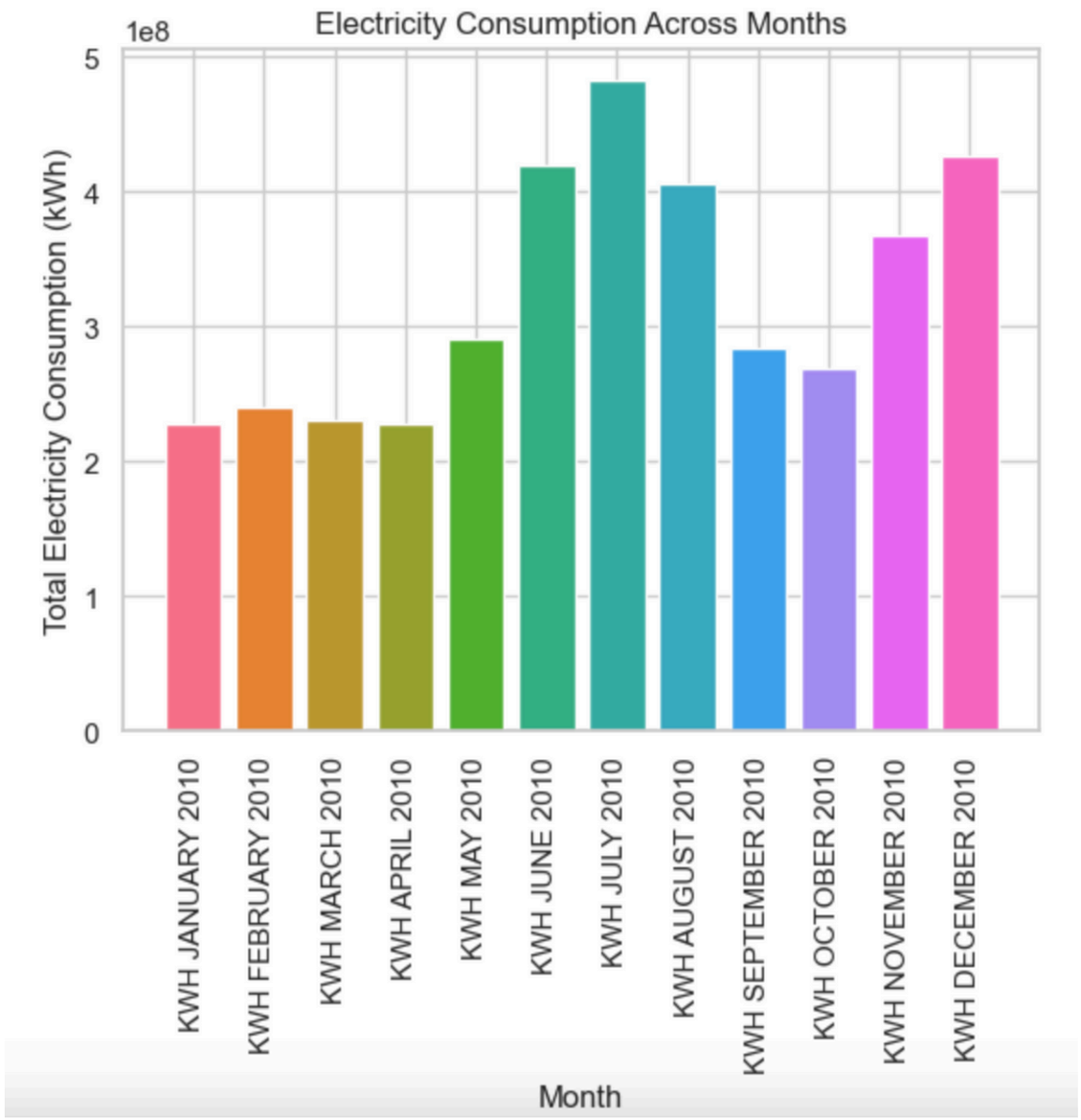


Çoklu < 7: 27284
Tek Aile: 26366
Ticari: 7775
Çoklu 7+: 5225
Belediye: 282
Endüstriyel: 42

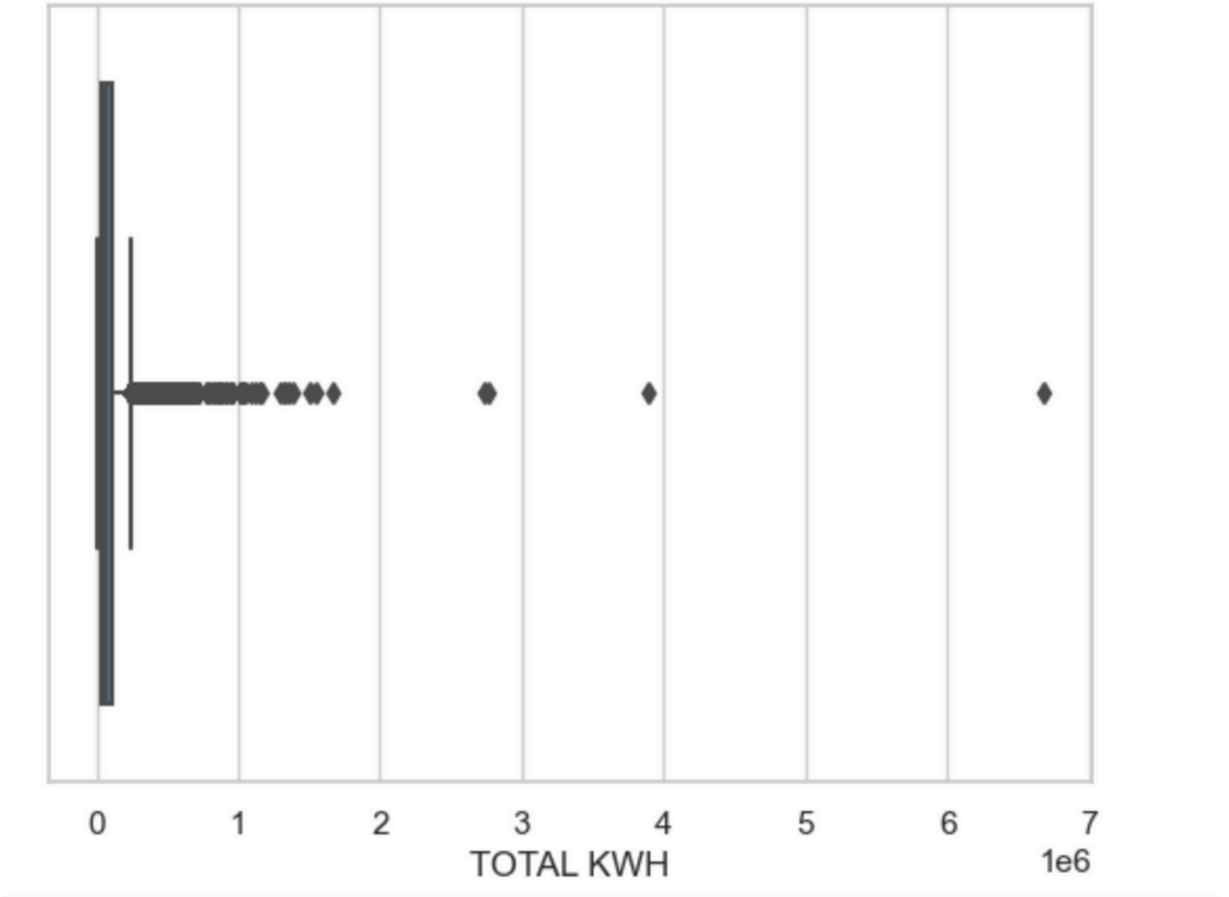
Veri Analizi ve Temizleme: Veri İnceleme



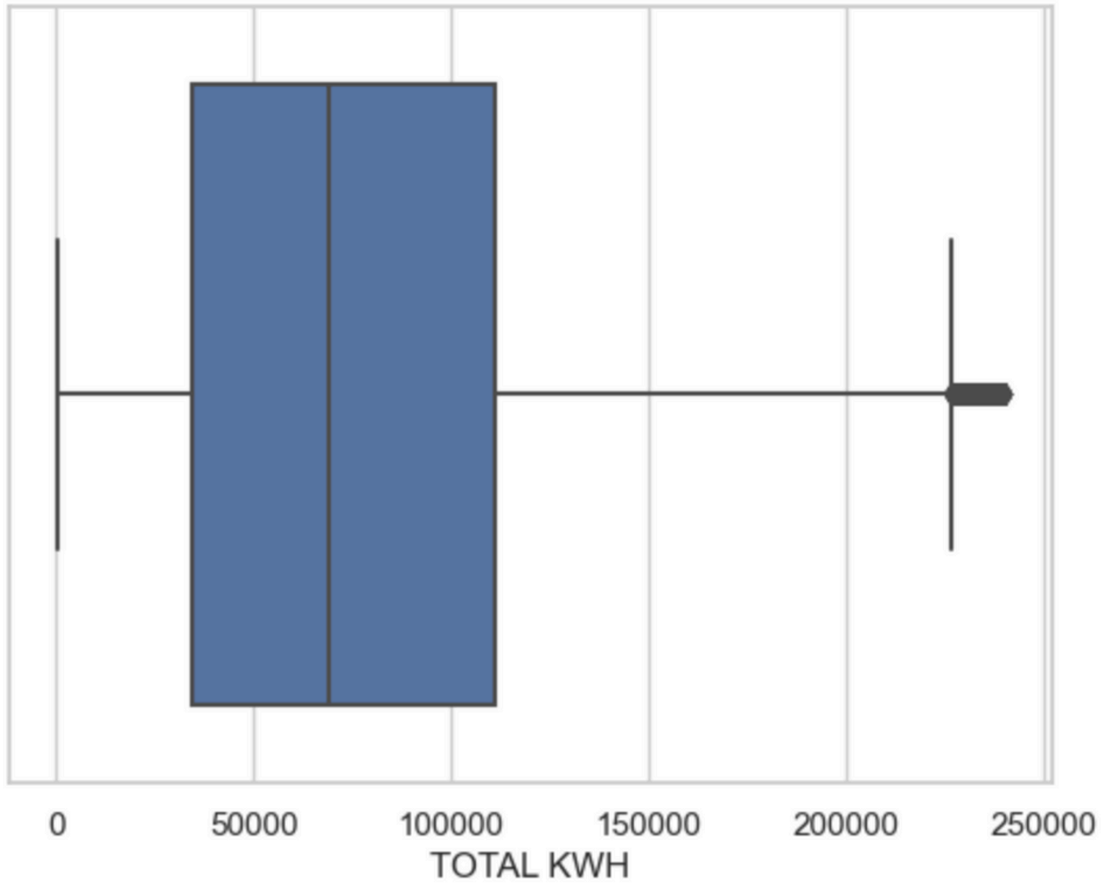
Veri Analizi ve Temizleme: Veri İnceleme



Veri Analizi ve Temizleme: Aykırı Değerlerin Tespiti

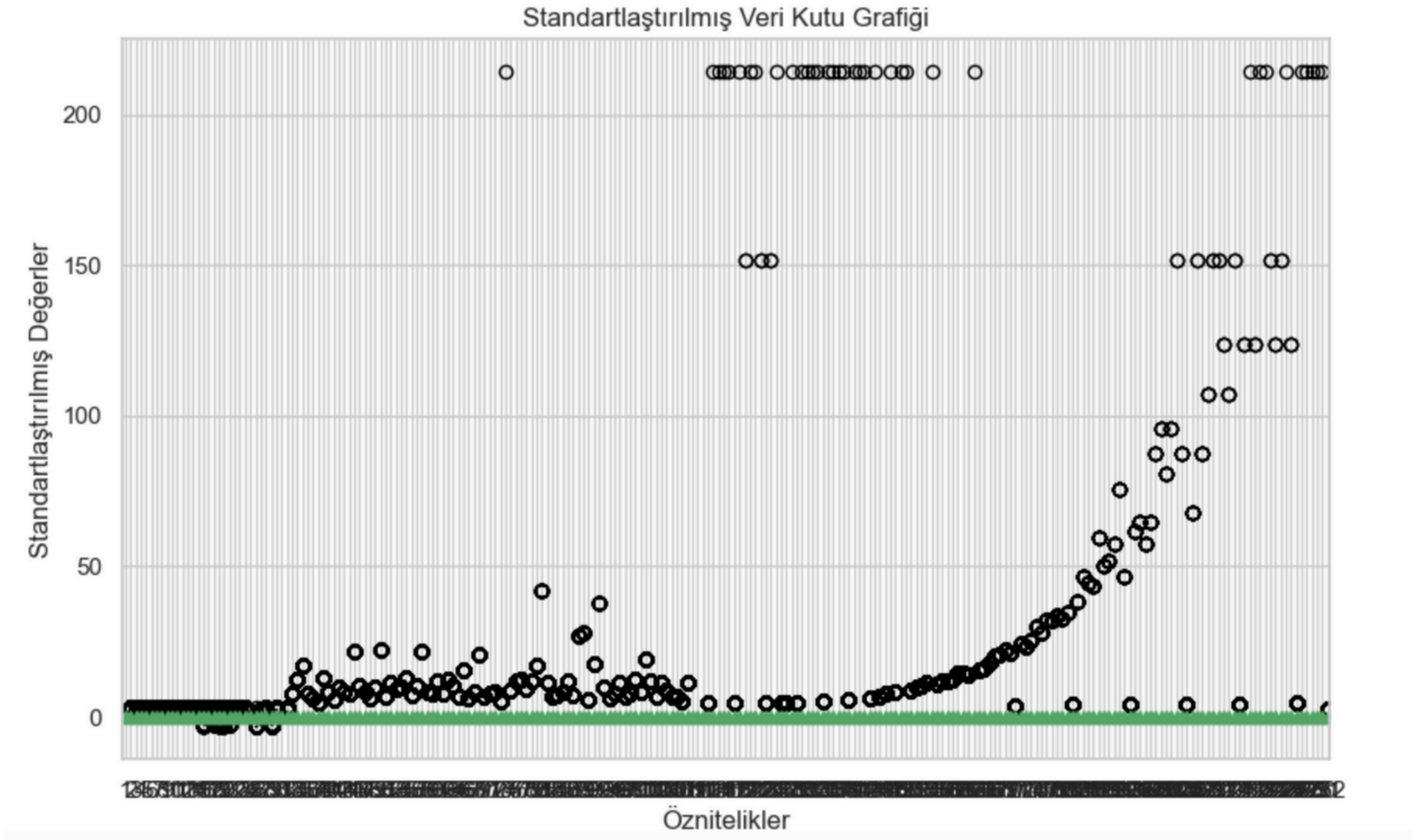


Veri Analizi ve Temizleme: Aykırı Değerlerin Düzeltilmesi



TOTAL KWH: 2010'da Tüketilen toplam KWH	**BUILDING SUBTYPE :** Yapıların kişi sayısı	RENTER-OCCUPIED HOUSING UNITS: Kiracıların kullandığı konut sayısı	KWH MAY 2010: Mayıs ayı harcanan toplam KWH
KWH NOVEMBER 2010: Kasım ayı harcanan toplam KWH	**KWH DECEMBER 2010:** Aralık ayı harcanan toplam KWH	ZERO KWH ACCOUNTS: 0 Enerji harcama miktarı	KWH MEAN 2010: Ort harcanan KHW
ELECTRICITY ACCOUNTS: Elektrik kullananların sayısı	**TOTAL UNITS :** Evlerin toplam sayısı	**RENTER-OCCUPIED HOUSING PERCENTAGE:** Kiracıların kullandığı konut yüzdesi	**KWH JUNE 2010:** Haziran ayı harcanan toplam KWH
TOTAL POPULATION: Toplam nüfus	**AVERAGE STORIES:** Yapıların ort kat sayısı	**KWH JANUARY 2010:** Ocak ayı harcanan toplam KWH	**KWH JULY 2010:** Temmuz ayı harcanan toplam KWH
OCCUPIED UNITS: Kullanılan toplam konut sayısı	**AVERAGE BUILDING AGE:** Yapıların ort yaşı	**KWH FEBRUARY 2010:** Şubat ayı harcanan toplam KWH	**KWH AUGUST 2010:** Ağustos ayı harcanan toplam KWH
COMMUNITY AREA NAME : Topluluk alanlarının ismi	**AVERAGE HOUSESIZE:** Evlerin ort yaşı	**KWH MARCH 2010:** Mart ayı harcanan toplam KWH	**KWH SEPTEMBER 2010:** Eylül ayı harcanan toplam KWH
BUILDING TYPE : Yapı türleri	**OCCUPIED UNITS PERCENTAGE :** Kullanılan konut sayıis yüzdesi	**KWH APRIL 2010:** Nisan ayı harcanan toplam KWH	**KWH OCTOBER 2010:** Ekim ayı harcanan toplam KWH

Veri Analizi ve Temizleme: Standartizasyon



Veri Seti Bölünmesi

Bağımlı Değişkenler: “TOTAL KWH” dışındaki tüm özellikler
Bağımsız Değişken: “TOTAL KWH”

Train-Test-Split: %25 test seti, %75 eğitim seti

Makine Öğrenimi Modelleri

LinearRegression

DecisionTreeRegressor

MultiOutputRegressor

XGBRegressor

GradientBoostingRegressor

Lineer Regresyon

Bağımlı değişkenin (toplam elektrik harcaması) bağımsız değişkenlerle (aylık elektrik tüketimleri, nüfus, bina özellikleri vb.) doğrusal bir ilişkisi olduğu düşünülerek Lineer Regresyon modeli denenmiştir.

Aşağıda eğitim ve test tahmin sonuçları verilmiştir.

	Veri Seti	MAE	MSE	R-squared
0	Eğitim	7.779849e-02	4.638366e-02	9.533489e-01
1	Test	5.201175e+09	6.701020e+22	-6.590392e+22

Decision Tree Regressor

Karmaşık kalıpları işleme konusundaki etkinlikleriyle bilinen karar ağacı modellerini kullanarak tahmin doğruluğunu ve sağlamlığını geliştirmek amaçlanmıştır.

Süreç, potansiyel aşırı uyumun ve azaltılmış genelleme yeteneğinin sinyalini veren yüksek varyans sorununu ortaya çıkarmıştır.

	Veri Seti	R-squared	MAE	MSE
0	Eğitim	1.000000	0.000000	0.000000
1	Test	0.912705	0.117602	0.088761

Multi Output Regressor

Karar ağacı regresyon modeli kullanarak yapılan tahminlerin sonuçlarına göre, model eğitim setinde mükemmel bir uyum göstermiştir (R-kare=1.0).

Test setinde de yüksek bir R-kare değeri elde edilmiştir (R-kare=0.9132), ancak eğitim setine kıyasla biraz daha düşüktür, bu da modelin hafif bir overfitting eğiliminde olduğunu göstermektedir.

	Metric	KWH Train Set	KWH Test Set
0	R-squared	1.0	0.9132
1	MAE	0.0	0.1179
2	MSE	0.0	0.0883

Gradient Boosting

Gradient Boosting algoritması, eğitim ve test setleri üzerinde farklı performans sergilemiştir.

Eğitim setindeki düşük hata oranları (MSE ve MAE) ve yüksek R^2 değeri (0.9656) modelin eğitim verilerine çok iyi uymuş olduğunu göstermektedir.

Test setinde bu performans önemsenmeyecek miktarda bir düşüş göstermiştir.

	Set	MSE	MAE	R^2
0	GB Eğitim	0.034169	0.078060	0.965634
1	GB Test	0.045491	0.085649	0.955260

XGBOOST

XGBoost Regressor modeli kullanılarak yapılan tahminlerin sonuçlarına göre, model eğitim setinde yüksek bir performans sergilemiştir.

Ortalama Mutlak Hata (MAE) değeri eğitim setinde 0.0376 iken, ortalama karesel hata (MSE) değeri 0.0032 olarak hesaplanmıştır.

Ayrıca, eğitim setindeki R-kare (R^2) değeri oldukça yüksektir ve %99.68 oranında varyansı açıklamaktadır.

Test seti için ise MAE değeri 0.0702, MSE değeri ise 0.0376 olarak hesaplanmıştır.

Test seti R-kare değeri ise %96.30 olarak bulunmuştur.

Bu sonuçlar, modelin eğitim setindeki performansını test setinde de büyük ölçüde koruduğunu göstermektedir, bu da modelin güvenilir ve tutarlı tahminler yaptığını göstermektedir.

	Veri Seti	MAE	MSE	R^2
0	Eğitim	0.037606	0.003223	0.996758
1	Test	0.070231	0.037580	0.963041

Modeller	Test R2 Skoru	Test MSE Skoru	Test MAE Skoru
Linear Regression	-6.590392+22	6.701020e+22	5.201175e+09
Decision Tree Regressor	0.912705	0.088761	0.117602
Multi Output Regressor	0.9132	0.0883	0.0
XGB Regressor	0.963041	0.037580	0.070231
Gradient Boosting Regressor	0.955260	0.045491	0.085649

Model Değerlendirmesi ve Sonuç

Lineer Regresyon modeli, modelin aşırı uyum (overfitting) yaşadığını ve eğitim setine fazlasıyla bağımlı hale geldiğini düşündüren sonuçlarla karşılaşılmıştır. Model, eğitim setindeki gürültüyü ve dalgalanmayı ezberlemiş olabilir ve yeni verilere uygulandığında başarısız olabilir.

DecisionTreeRegressor modeli, eğitim setinde mükemmel bir şekilde uymuş gibi görünüyor, ancak test setinde R-kare değeri düşük ve MSE yüksektir, bu da modelin aşırı uyum gösterdiğini ve genelleme yapamadığını gösterir.

MultiOutputRegressor modeli, modelinin eğitim setinde mükemmel bir şekilde performans gösterdiği, ancak test setinde bir miktar hata yaparak tahminlerinde belirli bir hata marjı olduğu görülmektedir. Bu sonuçlar, modelin genel olarak iyi performans gösterdiğini ancak belirli bir hata marjı olduğunu göstermektedir.

XGBRegressor modeli, eğitim ve test setlerinde oldukça iyi performans göstermektedir. R-kare değeri yüksek ve MSE düşüktür, bu da modelin verileri iyi açıkladığını ve tahminlerin gerçek değerlerle uyumlu olduğunu gösterir.

GradientBoostingRegressor modeli, eğitim ve test setlerinde yüksek R-kare değerleri ve düşük MSE değerleri ile oldukça iyi performans sergilemektedir. Bu da modelin verilere çok iyi uyum sağladığını ve tahminlerinin gerçek değerlere yakın olduğunu gösterir.

Model Değerlendirmesi ve Sonuç

Sonuç olarak, XGBOOST modelinin, bu veri seti üzerinde en iyi performansı gösterdiğini söyleyebiliriz. Bu model, diğer modellere kıyasla daha düşük MSE ve daha yüksek R-kare değerleri ile daha iyi tahminler yapmaktadır.

```
Ocak ayı elektrik harcaması: 1462393.2
Şubat ayı elektrik harcaması: 1233222.1
Mart ayı elektrik harcaması: 1126425.2
Nisan ayı elektrik harcaması: 1189878.0
Mayıs ayı elektrik harcaması: 1138147.2
Haziran ayı elektrik harcaması: 1066637.9
Temmuz ayı elektrik harcaması: 1356044.2
Ağustos ayı elektrik harcaması: 1256312.6
Eylül ayı elektrik harcaması: 1256390.0
Ekim ayı elektrik harcaması: 1368428.0
Kasım ayı elektrik harcaması: 1073654.8
Aralık ayı elektrik harcaması: 1168019.4
Toplam elektrik harcaması: 14695553.0
```

Teşekkürler :)