Rapor: Veri Analizi ve Manipülasyonu

Görev 1: Veri Temizleme ve Manipülasyonu

Keşifçi Veri Analizi adımları aşağıdaki sırayla yapılmıştır. Task'daki görevler burada mevcuttur.

1. Veri Setinin Genel Özeti
Toplam Gözlem Sayısı: 5000
Toplam Değişken Sayısı: 13 (Unnamed: 0 sütunu temizlenerek değişken sayısı 14'ten 13'e indirildi)
Değişken Türleri:
Kategorik: 5
Sayısal: 7
Kategorik Ancak Kardinal: 1

---

2. Eksik Veri Analizi
Eksik veri bulunan sütunlar:
FIYAT: 42 eksik değer (%0.84)
TOPLAM_SATIS: 4 eksik değer (%0.08)

---

3. Eksik Değerlerin Doldurulması
TOPLAM_SATIS Sütunu:
FIYAT ve ADET değerleri mevcut olan satırlarda, TOPLAM_SATIS = FIYAT x ADET formülüyle doldurulduu.
FIYAT Sütunu:
TOPLAM_SATIS ve ADET bilgisi mevcut olan satırlarda, FIYAT = TOPLAM_SATIS / ADET formülüyle doldurulduu.
Eksik değerlerin doldurulamayan kısmı, sütunların medyan değerleriyle tamamlandı.

---

4. Aykırı Değerlerin İncelenmesi
Aykırı değer kontrolü sonuçları:
TOPLAM_SATIS: Aykırı değer tespit edildi.
Diğer sayısal değişkenlerde (örneğin, YAS, HARCAMA_MIKTARI, FIYAT, ADET) aykırı değer bulunmadı.
Aykırı Değerlerin İşlenmesi:
Aykırı değerler, eşik değerlerle baskılandı (alt ve üst sınırlarla değiştirildi).

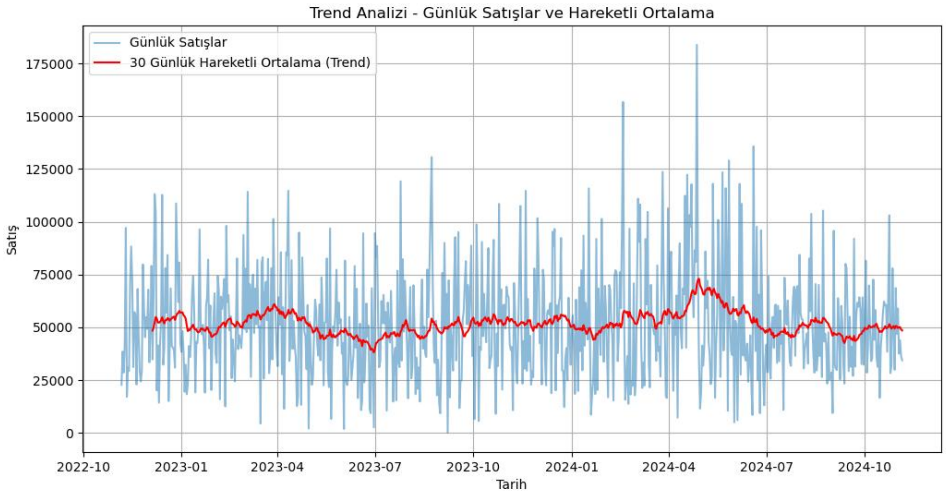
Görev 2: Zaman Serisi Analizi

Zaman Serisi görevlerine başlamadan önce zaman serisi analizi yapılmıştır. Burada trend, mevsimsellik ve durağanlık analiz edilmiştir.

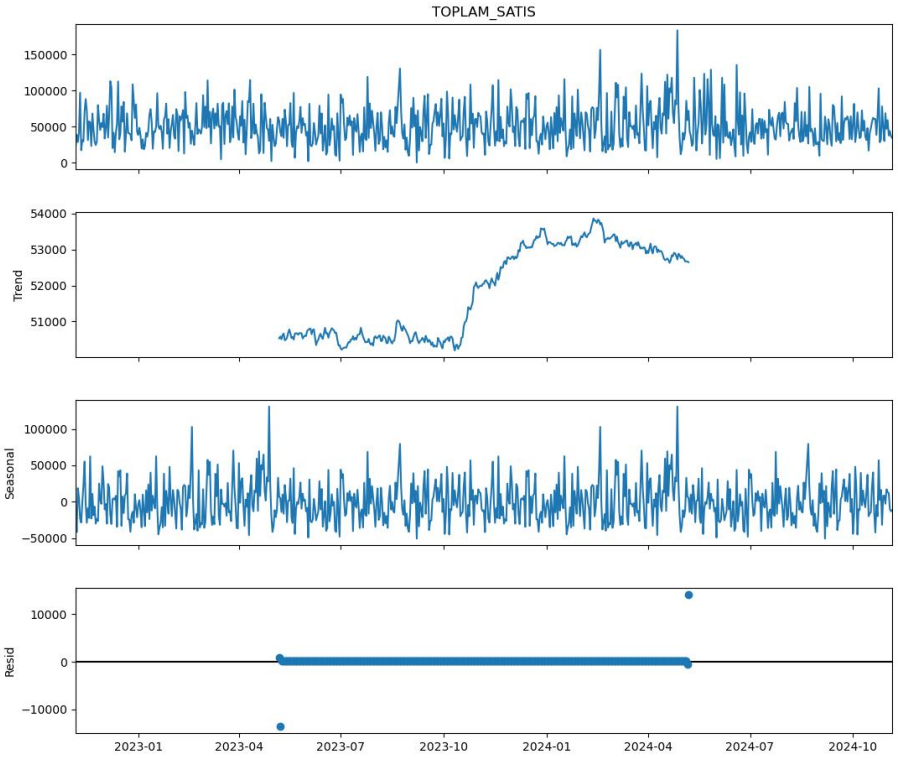
Veri Hazırlığı
Zaman Serisi Verisi: Verinin tarihsel bileşeni olan TARIH sütunu, indeks olarak ayarlandı.
Günlük Toplam Satışlar: Veri, günlük toplam satışları analiz etmek için resample('D') fonksiyonu ile yeniden örneklenip toplandı.

Analiz Adımları

a) Trend Analizi
Grafik: Zaman içindeki trendin gözlemlenebilmesi için bir grafik oluşturuldu.



b) Mevsimsellik Analizi
Grafik: Mevsimsel bileşenlerin belirlenebilmesi amacıyla bir mevsimsellik analizi gerçekleştirildi.
Yorum: Veri mevsimsel bir örüntü göstermektedir.



c) Durağanlık Analizi
ADF (Augmented Dickey-Fuller) Testi: Zaman serisinin durağan olup olmadığını test etmek için ADF testi uygulandı.
Test İstatistiği: -26.06
p-Değeri: 0.0
Kritik Değerler:
%1: -3.44
%5: -2.87

%10: -2.57
Sonuç: Test istatistiği kritik değeri aşmakta ve p-değeri 0.05'ten küçük olduğundan, zaman serisi durağandır.

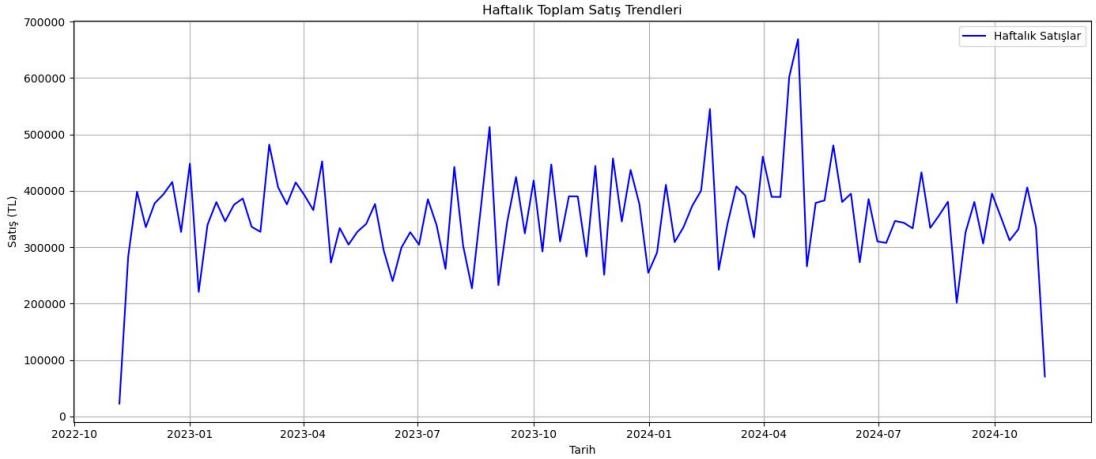
Sonuçlar:
Durağanlık: Zaman serisi durağandır çünkü ADF testinden elde edilen p-değeri (0.0) 0.05'ten küçük.
Sonuç: Zaman serisi durağan olup, trend ve mevsimsellik gibi bileşenler gözlemlenmiştir.

ADIMLAR:

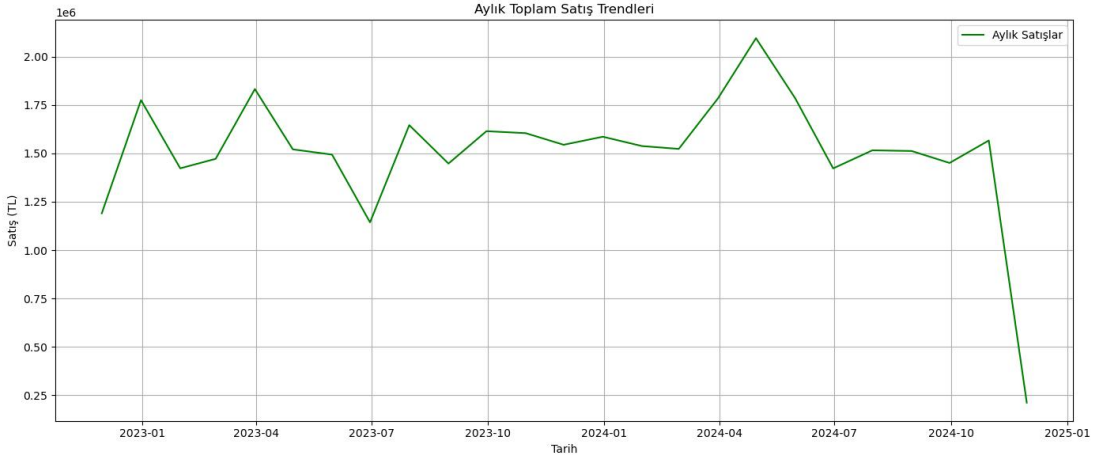
1. Haftalık ve Aylık Satış Analizi

Haftalık ve Aylık Toplam Satış Grafikleri:

Haftalık Satışlar: Haftalık toplam satışları incelemek için total_weekly_sales(data) fonksiyonu çağrıldı ve grafikler oluşturuldu.

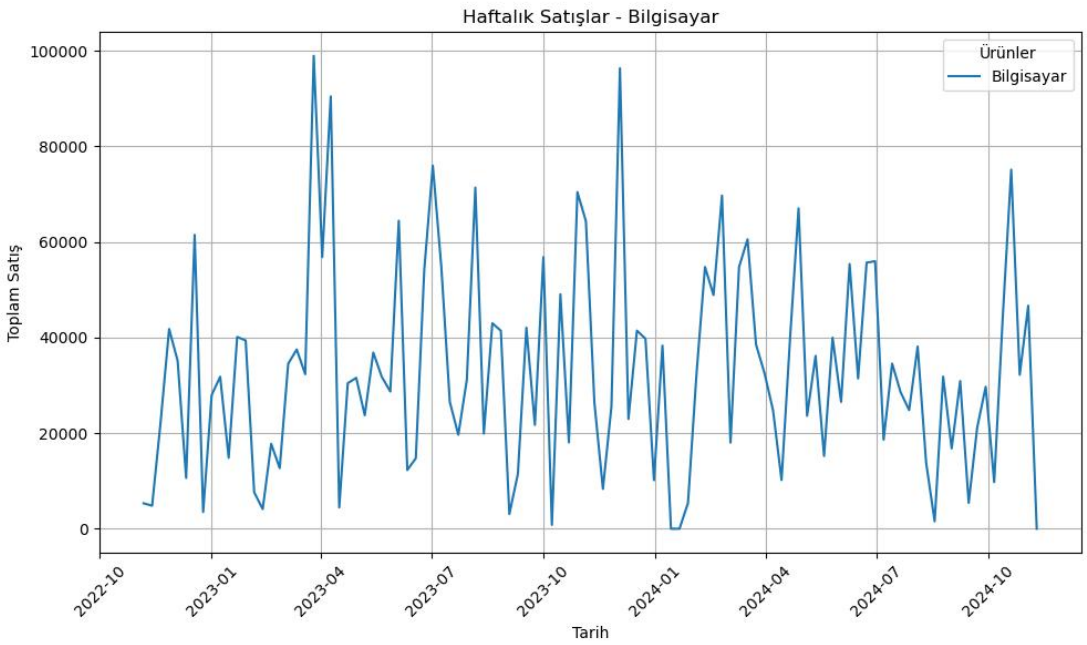


Aylık Satışlar: Aylık toplam satışlar için total_monthly_sales(data) fonksiyonu ile grafikler elde edildi.

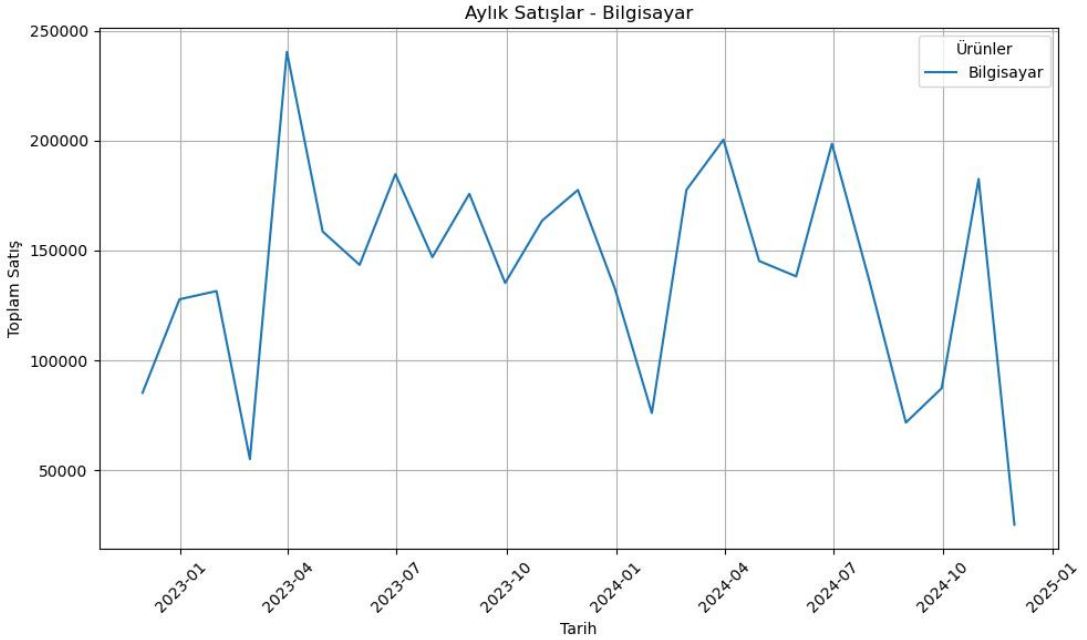


Haftalık ve Aylık Ürün Satışları:

Haftalık Ürün Satışları: Her bir ürün için haftalık satışlar weekly_sales_trends(data) fonksiyonu ile analiz edildi ve grafikler çizildi.



Aylık Ürün Satışları: Aylık bazda ürün satış trendleri için monthly_sales_trends(data) fonksiyonu kullanıldı ve ürünlerin satış hacimleri görselleştirildi.



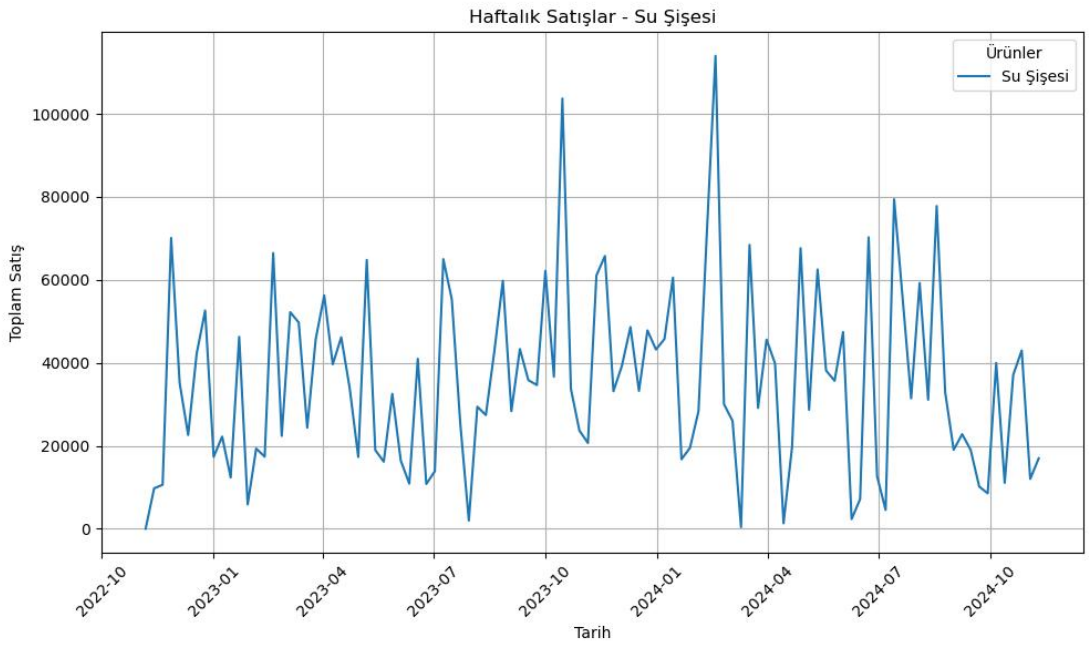
---

2. Ayın İlk ve Son Satış Günleri

Her ayın ilk ve son günleri bulunmuş ve hesaplanarak kod betiğine eklenmiştir.

---

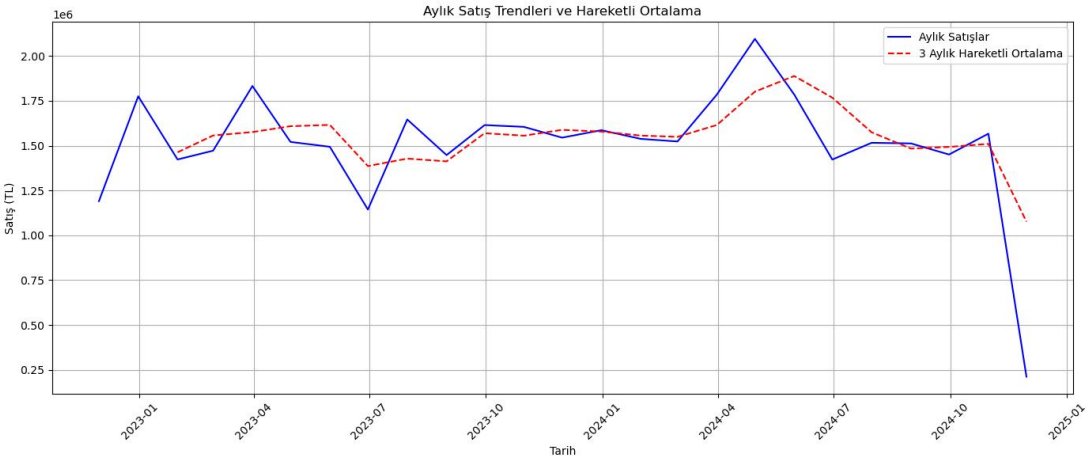
3. Haftalık Ürün Adedi
Haftalık Ürün Adedi (Ürün Bazında)
Ürün bazında haftalık satış adedi hesaplanarak her bir ürün için haftalık satış trendleri elde edildi:



Haftalık Ürün Adedi (Toplam)
Haftalık toplam ürün adedi de hesaplanarak her hafta için genel satış adedi analiz edildi:

---

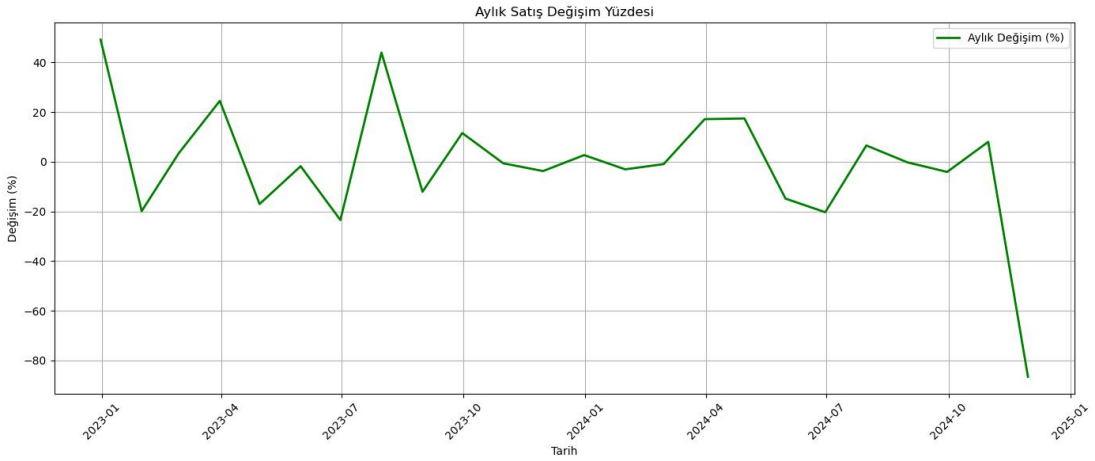
4. Aylık Satışlar ve Hareketli Ortalama
Aylık Satışlar ve Hareketli Ortalama: Aylık satış verilerinin hareketli ortalamaları monthly_sales_averages(data) fonksiyonu ile hesaplandı ve grafikler elde edildi.



---

5. Aylık Satış Değişim Yüzdesi
Aylık Satış Değişim Yüzdesi: Aylık satış değişim yüzdesi hesaplanarak verilerdeki eğilimler analiz edildi:
python
Kodu kopyala
Aylık satış değişim yüzdesi

monthly_sales_change(data)



Grafikler ve Analizler: Haftalık ve aylık bazda satış eğilimleri incelenmiş, mevsimsel değişimler, hareketli ortalamalar ve değişim yüzdeleri gibi göstergelerle verilerin derinlemesine analizi yapılmıştır.

Görev 3: Kategorisel ve Sayısal Analiz

1. Ürün Kategorilerine Göre Toplam Satış ve Oran Analizi
2.
Toplam Satış Miktarı
Her bir ürün kategorisi için toplam satış miktarı hesaplandı ve satışlar içindeki oranları belirlendi.

KATEGORI	Toplam Satış	Satış Adedi	Satış Oranı (%)
Elektronik	18,262,440.00	2440	48.44%
Ev Aletleri	3,791,499.00	479	10.06%
Giyim	3,894,336.00	509	10.33%
Kırtasiye	7,986,450.00	1056	21.18%
Mutfak Ürünleri	3,766,484.00	516	9.99%

Ürün Bazında Satışlar
Her bir ürünün toplam satış miktarı, satış adedi ve oranları aşağıda verilmiştir:

ÜRÜN_ADI	Toplam Satış	Satış Adedi	Satış Oranı (%)
Bilgisayar	3,498,532.00	449	9.28%
Defter	3,966,806.00	523	10.52%
Fırın	3,791,499.00	479	10.06%
Kalem	4,019,644.00	533	10.66%
Klima	3,355,930.00	487	8.90%
Kulaklık	3,661,305.00	482	9.71%
Mouse	3,802,867.00	499	10.09%
Su Şişesi	3,766,484.00	516	9.99%

ÜRÜN_ADI	Toplam Satış	Satış Adedi	Satış Oranı (%)
Telefon	3,943,807.00	523	10.46%
Çanta	3,894,336.00	509	10.33%

---

2. Yaş Gruplarına Göre Satış Eğilimleri
Yaş Kategorisi Oluşturma
Müşterilerin yaşları şu şekilde 4 gruba ayrıldı:
18-25
26-35
36-50
50+
Yaş Gruplarına Göre Satış Ortalamaları
Her yaş grubunun ortalama harcama miktarları aşağıda verilmiştir:

Yaş Grubu	Ortalama Satış
18-25	8,009.82
26-35	7,399.40
36-50	7,526.36
50+	7,428.05

Bu veriler, farklı yaş gruplarındaki müşterilerin harcama miktarlarının ortalama olarak oldukça yakın olduğunu göstermektedir.

---

3. Kadın ve Erkek Müşterilerin Harcama Miktarları
Kadın ve Erkek Müşteriler Arasındaki Harcama Farkları

Cinsiyet	Ortalama Harcama
Erkek	2,571.32
Kadın	2,610.13

Kadın ve erkek müşterilerin harcama miktarları arasında küçük bir fark bulunmaktadır; kadınlar ortalama olarak erkeklerden daha fazla harcama yapmaktadır.

Sonuçlar
Bu analiz, ürün kategorilerine, yaş gruplarına ve cinsiyetlere göre satış eğilimlerini derinlemesine incelemekte olup, harcama davranışları ve satış trendleri hakkında bilgi sağlamaktadır. Elektronik ürünler en yüksek satış oranına sahipken, yaş grupları arasındaki farklar nispeten küçüktür. Kadınlar, erkeklere kıyasla biraz daha fazla harcama yapmaktadır.

Görev 4: İleri Düzey Veri Manipülasyonu

1. Şehir Bazında Toplam Harcama
Müşterilerin şehirlerine göre toplam harcama miktarları hesaplanıp, şehirler en çok harcama yapanlardan başlayarak sıralandı.

Şehir	Toplam Harcama (₺)
Gaziantep	1,740,073.60
İzmir	1,739,145.40
Bursa	1,665,417.27
Ankara	1,649,167.77
Antalya	1,581,436.54
Konya	1,556,038.91
Adana	1,552,729.00

Şehir	Toplam Harcama (₺)
İstanbul	1,470,821.34

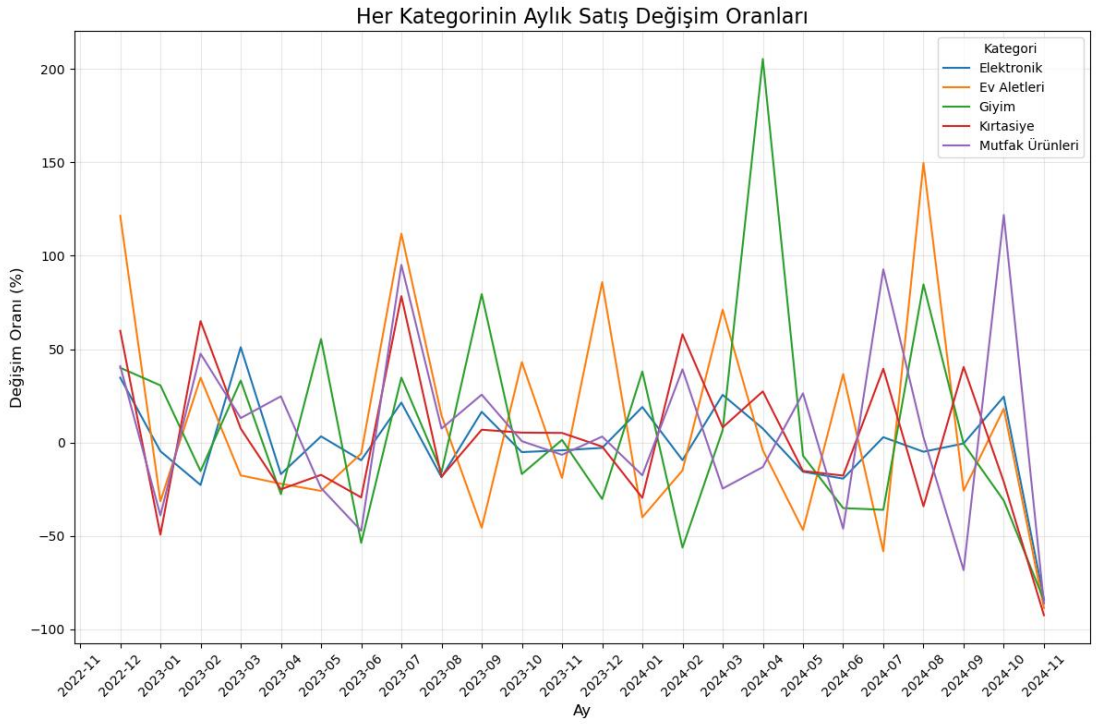
Bu sıralama, harcama miktarı en yüksek olan şehirlerin başta geldiğini göstermektedir.

---

2. Ürün Bazında Ortalama Satış Artışı
Her bir ürün için ortalama satış artışı oranı hesaplandı. Bu oran, her ürün için bir önceki aya göre satış değişim yüzdesine dayanmaktadır.

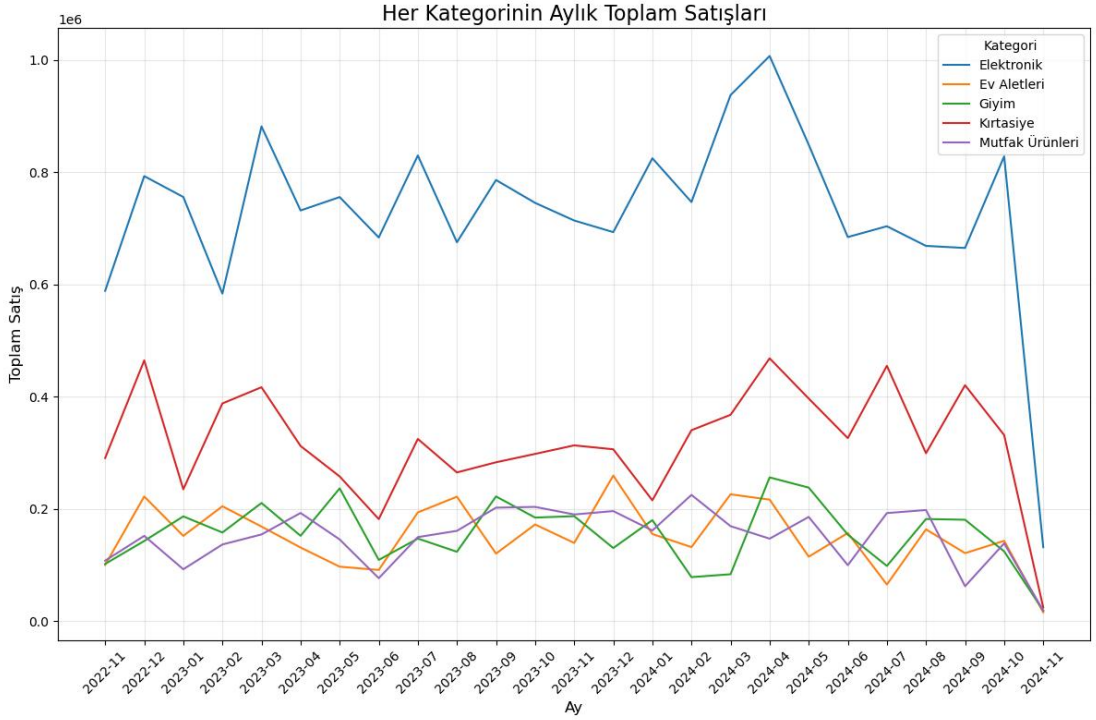
Ürün Adı	Ortalama Satış Artışı (%)
Bilgisayar	15.70%
Defter	7.76%
Fırın	9.99%
Kalem	9.56%
Klima	6.59%
Kulaklık	10.02%
Mouse	5.35%
Su Şişesi	6.98%
Telefon	0.46%
Çanta	8.27%

Bu oran, her bir ürünün satışlarının bir önceki aya göre ne kadar arttığını gösteriyor. Bilgisayar ürününde en yüksek artış oranı gözlemlenirken, Telefon ürününde artış çok düşük kalmıştır.



---

3. Kategori Bazında Aylık Toplam Satış ve Değişim
Pandas groupby fonksiyonu kullanılarak her bir ürün kategorisinin aylık toplam satışları hesaplandı ve bu satışlardaki değişim oranları grafikle
gösterildi.



Bu analiz, her kategorinin satış trendlerini görsel olarak incelemenizi sağlar. Grafik ile satışlar arasındaki aylık değişimleri daha net bir şekilde görebilirsiniz.

Sonuçlar
Şehir Bazında Harcama: Gaziantep ve İzmir en yüksek harcama yapan şehirler olarak öne çıkmaktadır.
Ürün Bazında Satış Artışı: Bilgisayar ve Kulaklık gibi ürünlerde yüksek satış artışı gözlemlenirken, Telefon gibi ürünlerde artış düşük kalmıştır.
Kategori Aylık Satış ve Değişim: Satışların aylık periyotlarda nasıl değiştiğini görselleştirerek, her kategorinin trendlerini daha iyi analiz edebilirsiniz.

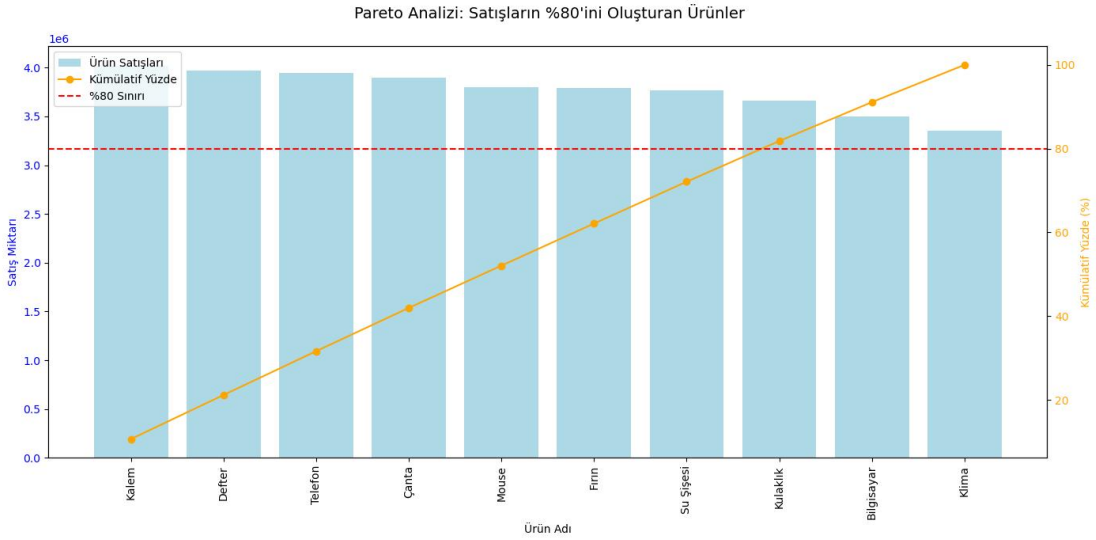
Görev 5: BONUS

Görev 5: Pareto / Cohort / Modelleme
1. Pareto Analizi
Pareto analizi, satışların %80'ini oluşturan ürünleri belirlemekte kullanıldı. Analiz sonucu, satışların büyük kısmını oluşturan ürünler aşağıdaki gibi sıralandı:

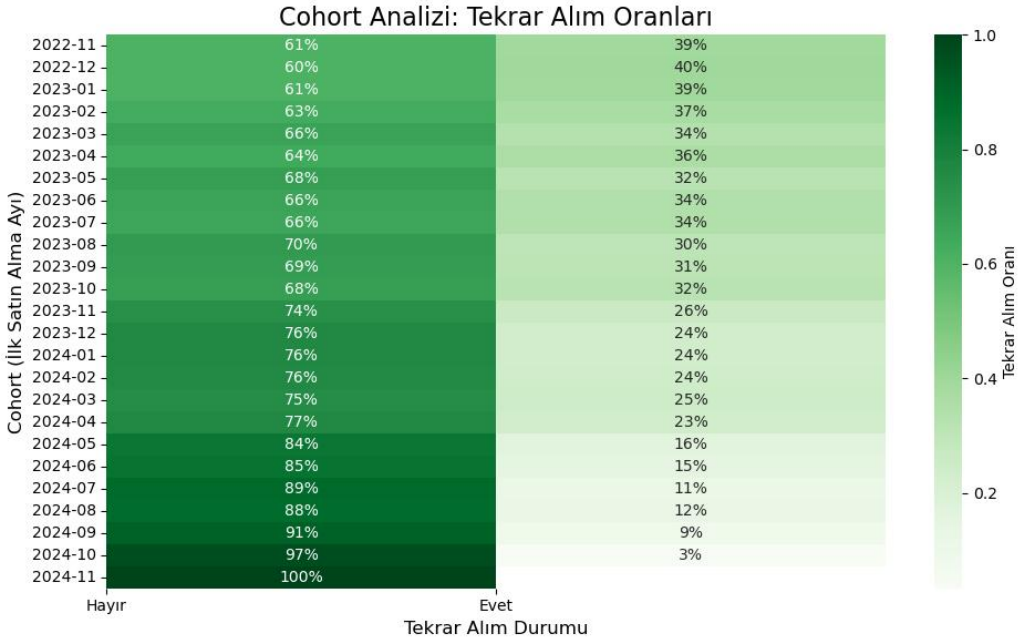
Satışların %80'ini Oluşturan Ürünler:

Kalem
Defter
Telefon
Çanta
Mouse
Fırın
Su Şişesi

Pareto grafiği, satış miktarlarının ve kümülatif yüzdelerin görselleştirilmesi ile satışları yönlendiren en önemli ürünleri gösterdi.



2. Cohort Analizi
Cohort analizi, müşteri bazında tekrar alım oranlarını inceledi. İlk satın alma tarihine göre yapılan tekrar alımlar analiz edilerek, her cohort (ilk satın alma dönemi) için tekrar alım oranları hesaplandı ve görselleştirildi.



Cohort Analizi: Tekrar Alım Oranları (Isı Haritası):
Evet: Tekrar alım yapan müşterilerin oranı
Hayır: Tekrar alım yapmayan müşterilerin oranı

Isı haritası, cohort'lar bazında tekrar alım oranlarını net bir şekilde gösterdi. Bu analiz, hangi dönemlerde müşterilerin tekrar alım yapma oranlarının daha yüksek olduğunu ve hangi cohort'larda bu oranların daha düşük olduğunu görmeyi sağladı.

Cohort analizinin sonuçları, her cohort (ilk satın alma dönemi) için tekrar alım oranlarını göstermektedir. Burada her bir satır, belirli bir ayda oluşturulmuş bir cohort ve o cohort'un

tekrar alım yapıp yapmadığına dair oranları içermektedir. Verilerde True ve False değerleri, müşterilerin tekrar alım yapma durumlarını (yani ikinci bir alışveriş yapıp yapmadıklarını) gösteriyor.

Örneğin:
2022-11 Cohort'u: Bu cohort'tan olan müşterilerin %39.37'si tekrar alım yapmışken, %60.63'ü tekrar alım yapmamış.
2024-11 Cohort'u: Bu cohort'tan müşterilerin tamamı (100%) ilk satın almayı yapmış, ancak henüz tekrar alım yapılmamış. Bu, henüz analiz edilen dönemin sonu olduğu için bir "NaN" değeri (veri eksikliği) gösteriyor.

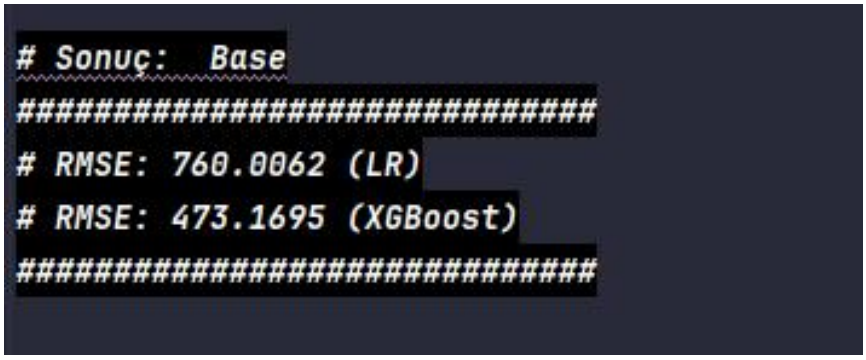
Sonuç Olarak:
Görsel olarak incelendiğinde, cohort'lar ilerledikçe tekrar alım oranlarında bir azalma gözlemleniyor. Özellikle 2024'ün başından itibaren, tekrar alım oranlarında önemli bir düşüş yaşanmakta, bu da kullanıcıların tekrar alışveriş yapma oranlarının giderek azaldığını gösteriyor. İlk dönemdeki cohort'larda tekrar alım oranı daha yüksekken, sonraki aylarda bu oran giderek düşmekte.
2024 yılının ilerleyen aylarında tekrar alım oranları hızla azalmış ve 2024-11 itibariyle hiç tekrar alım yapılmamış gibi görünüyor.

1. Modelleme

A)Veri Temizliği ve Feature Engineering:

Verisetindeki gereksiz sütunlar (ISIM, ÜRÜN_KODU, AY, ILK_SATINALMA, COHORT, SATINALMA_DONEMI, HARCAMA_MIKTARI) çıkarıldı.
Yeni değişkenler eklendi:
Haftanın günü, aylık ortalama fiyat, fiyat/adet oranı, ürün fiyatı ve satış miktarının çarpımı gibi özellikler türetildi.
Mevsimsel satış durumu (Kış, İlkbahar, Yaz, Sonbahar) oluşturuldu.
Yılbaşına kalan gün sayısı eklendi.
Boolean sütunları 0 ve 1'e dönüştürüldü.
Kategorik veriler için One-Hot Encoding ve Label Encoding uygulandı.

B) Model Seçimi ve Değerlendirmesi:
Modeller arasında Linear Regression (LR) ve XGBoost (gelişmiş regresyon) kullanıldı.
RMSE (Root Mean Squared Error) ile model performansı değerlendirildi:
Linear Regression: 760.0062 RMSE
XGBoost: 473.1695 RMSE
XGBoost, lineer regresyona göre daha iyi performans gösterdi. Bu, modelin doğruluğunun arttığını ve hata oranının daha düşük olduğunu gösteriyor.



C) Hiperparametre Optimizasyonu:
XGBoost modelinde GridSearchCV kullanılarak en iyi parametreler bulunmaya çalışıldı.
Optimizasyon sonucunda modelin performansı iyileştirildi:
RMSE: 447.2684
Hiperparametre optimizasyonu sayesinde modelin doğruluğu önemli ölçüde arttı.
Sonuç:
Başlangıçta kullanılan XGBoost modeli, hiperparametre optimizasyonu ile önemli bir iyileşme gösterdi.
İyileştirilmiş modelin performansı, test seti üzerinde daha düşük bir hata ile sonuçlanmıştır.