PROJE DURUMU RAPOR

Proje Adı	Veri Seti Analizi	**RAPOR PERİYODU**
Proje Sahibi	Kaan Balcı	05/12/2024 - 05/19/2024
Öğretmen	Burak Evrentuğ	

ÖNE ÇIKANLAR

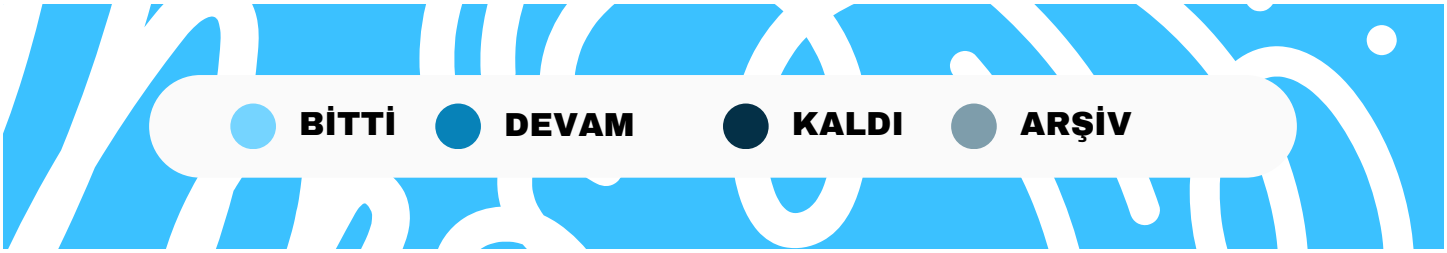
- Gerçek dünya verileri üzerinde veri ön işleme tekniklerini uygulayarak ve uygun makine öğrenmesi modelleri geliştirerek öğrencilerin veri bilimi ve makine öğrenmesi konularındaki bilgi ve becerilerini pekiştirmektir.
- Öğrenciler, seçtikleri bir veri seti üzerinde analiz yapacak, veriyi temizleyecek, işleyecek ve son olarak veriden anlamlı sonuçlar çıkarmak için makine öğrenmesi modelleri eğitecektir

REKABETLER

- **Proje Raporu:** Araştırma sorusunu, kullanılan veri setini, veri ön işleme adımlarını, seçilen makine öğrenmesi modelini/modellerini, model eğitim sürecini, model değerlendirme yöntemlerini ve elde edilen sonuçları açıklayan detaylı bir rapor.
- **Kod Dosyaları:** Veri ön işleme, model eğitimi ve değerlendirme adımlarını içeren, açıklamalarla zenginleştirilmiş Python notebook dosyası
- **Youtube Videosu:** Ortalama 5 dakikalık bir video ile yapılan çalışmaların açıklanması.
- **Ders Sunumu:** 10 dakikalık bir ders sunumu ile yapılan çalışmaların sınıf ortamında açıklanması.

DURUM GÜNCELLEME

GÖREV VEYA TESLİMAT	GÖREV SAHİBİ	DURUM
Veri Temizleme: Eksik değerleri doldurma veya kaldırma, aykırı değerleri işleme.	KAAN BALCI	BİTTİ
Veri Dönüşümü: Özellik ölçeklendirme, kategorik verilerin sayısal verilere dönüştürülmesi.	KAAN BALCI	BİTTİ
Veri Görselleştirme: Veri keşfi için çeşitli grafikler ve görseller kullanma.	KAAN BALCI	BİTTİ
Özellik Mühendisliği: Model performansını artırmak için yeni özellikler oluşturma ve seçme.	KAAN BALCI	BİTTİ



SONRAKİ ADIM

DEĞERLENDİRME KRİTERLERİ

GÖREV VEYA TESLİMAT	GÖREV SAHİBİ
Model Performansı: Sınıflandırma için: Doğruluk, hassasiyet, geri çağırma, F1 skoru Regresyon İçin: MAE, MSE ve RMSE gibi metriklerle model performansının değerlendirilmesi.	KAAN BALCI
Rapor ve Sunum: Proje raporunun açıklık ve bütünlüğü, analizlerin ve sonuçların net bir şekilde sunulması.	KAAN BALCI
Kod Kalitesi: Kodun okunabilirliği, açıklamaların varlığı ve kod organizasyonu.	KAAN BALCI

Veri seti

Veri seti, çeşitli araçların özellikleri ve satış fiyatları hakkında bilgiler içermektedir.

Amaç, aracın özelliklerine dayanarak satış fiyatını tahmin edebilecek bir makine öğrenimi modeli geliştirmektir. Bu tahmin modeli, potansiyel alıcıların ve satıcıların bir aracın adil piyasa değerini tahmin etmelerine yardımcı olacaktır.

1. **Car_ID**: Her araba listesinin benzersiz bir tanımlayıcısı.
2. **Brand**: Arabanın markası veya üreticisi (örneğin, Toyota, Honda, Ford, vb.).
3. **Model**: Arabanın modeli (örneğin, Camry, Civic, Mustang, vb.).
4. **Year**: Arabanın imalat yılı.
5. **Kilometers_Driven**: Araba tarafından toplam kilometre.
6. **Fuel_Type**: Arabanın kullandığı yakıt türü (örneğin, Benzin, Dizel, Elektrikli, vb.).
7. **Transmission**: Arabanın şanzıman tipi (örneğin, Manuel, Otomatik).
8. **Owner_Type**: Arabanın önceki sahiplerinin sayısı (örneğin, Birinci, İkinci, Üçüncü).
9. **Mileage**: Arabanın kilometre başına yakıt verimi.
10. **Engine**: Arabanın motor kapasitesi CC cinsinden (Santimetre Küp).
11. **Power**: Arabanın maksimum güç çıkışı beygir gücü (BHP) cinsinden.
12. **Seats**: Arabada bulunan koltuk sayısı.
13. **Price**: Arabanın INR (Hindistan Rupisi) cinsinden satış fiyatı, tahmin etmek için hedef değişken.

Veri seti Analizi

- Veri setinde 100 satır (row) bulunmaktadır.
- Arabaların ortalama kilometre kullanımı 28,150'dir.
- Arabaların ortalama yakıt verimi 17.21 kilometre/litre'dir.
- Arabaların ortalama motor kapasitesi 1855.23 CC'dir.
- Arabaların ortalama güç çıkışı 158.13 beygir gücüdür.
- Arabaların ortalama koltuk sayısı 5.23'tür.
- Arabaların ortalama satış fiyatı 1,574,000 INR'dir.
- **Car_ID:** 1 ile 100 arasında değişen benzersiz bir tanımlayıcıdır.
- **Year:** 2016 ile 2021 arasında değişen araçların imalat yıllarını gösterir.
- **Kilometers_Driven:** 15000 ile 60000 arasında değişen, ortalaması 28150 olan toplam kilometreleri gösterir.
- **Mileage:** 10 ile 25 arasında değişen, ortalaması 17.21 olan yakıt verimlilik değerlerini gösterir.
- **Engine:** 1197 ile 4951 arasında değişen, ortalaması 1855.23 olan motor kapasitesini CC cinsinden gösterir.
- **Power:** 74 ile 396 arasında değişen, ortalaması 158.13 olan maksimum güç çıkışını beygir gücü cinsinden gösterir.
- **Seats:** 4 ile 7 arasında değişen, ortalaması 5.23 olan koltuk sayısını gösterir.
- **Price:** 550000 ile 4000000 arasında değişen, ortalaması 1574000 olan araçların INR cinsinden satış fiyatlarını gösterir.
- Boşluk bulunmamaktadır.

Veri seti Aykırı Değer Analizi

Car_ID

Q1: 27.75 - Q3 : 75.75 - IQR : 48.0 - low : -44.25 - high : 147.75

Year

Q1: 2018.0 - Q3 : 2019.25 - IQR : 1.25 - low : 2016.125 - high : 2021.125

Kilometers_Driven

Q1: 22000.0 - Q3 : 30500.0 - IQR : 8500.0 - low : 9250.0 - high : 43250.0

Mileage

Q1: 17.0 - Q3 : 20.0 - IQR : 3.0 - low : 12.5 - high : 24.5

Engine

Q1: 1199.0 - Q3 : 1968.0 - IQR : 769.0 - low : 45.5 - high : 3121.5

Power

Q1: 94.0 - Q3 : 175.75 - IQR : 81.75 - low : -28.625 - high : 298.375

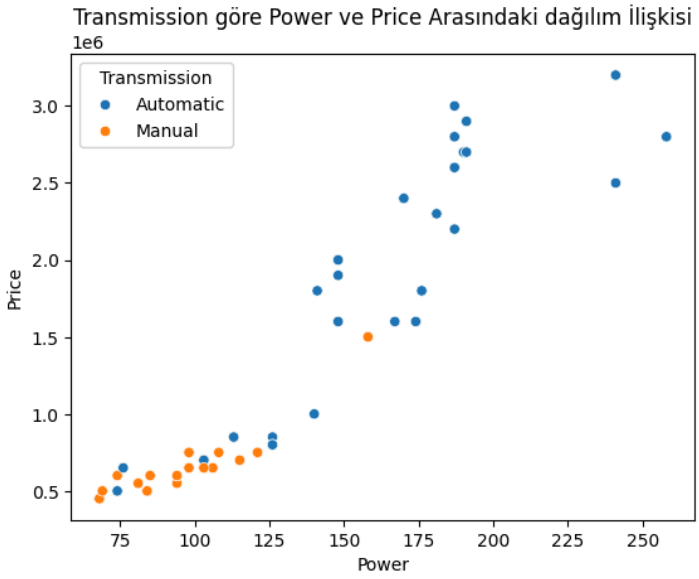
Seats

Q1: 5.0 - Q3 : 5.0 - IQR : 0.0 - low : 5.0 - high : 5.0

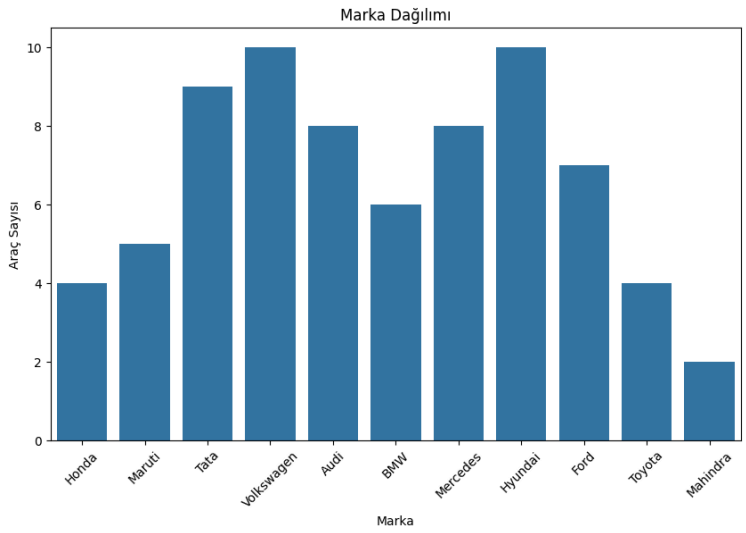
Price

Q1: 600000.0 - Q3 : 2325000.0 - IQR : 1725000.0 - low : -1987500.0 - high : 4912500.0

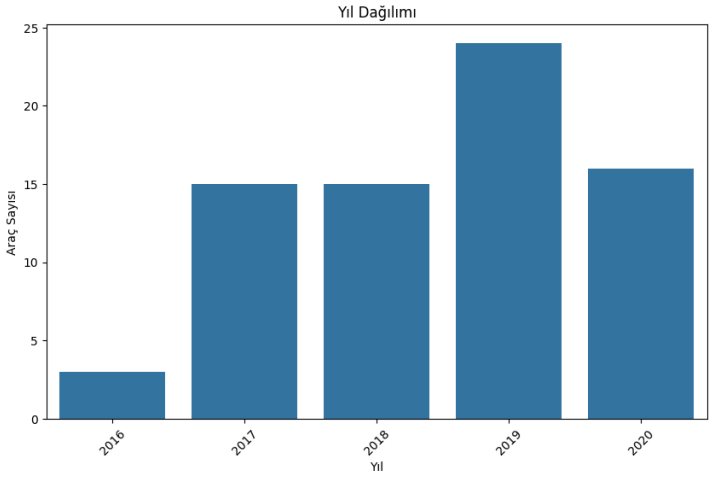
Veri seti Görselleştirme Analizi



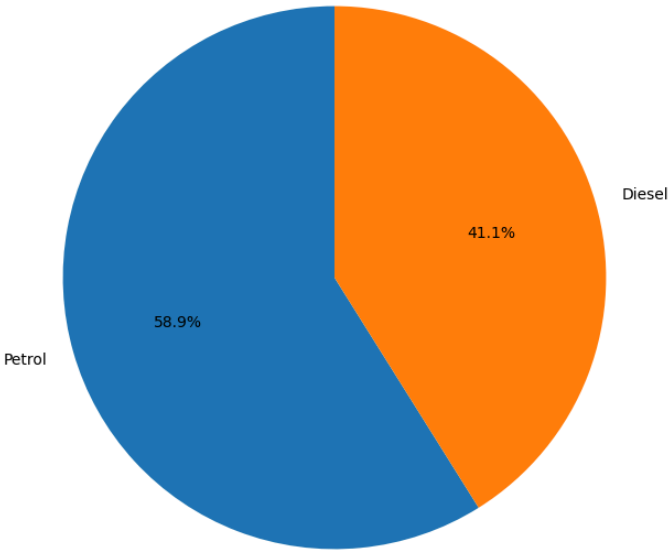
(Power) olarak max 250+ bir değerimiz var (fiyat) olarak max 3.000.000+ bir değerimiz var ve burada gördüğümüz üzere araç fiyatı arttıkça motor gücü doğru orantılı bir şekilde yükselmektedir ve (vites türünün) otomatik tercih edildiği analizini yapabiliriz.



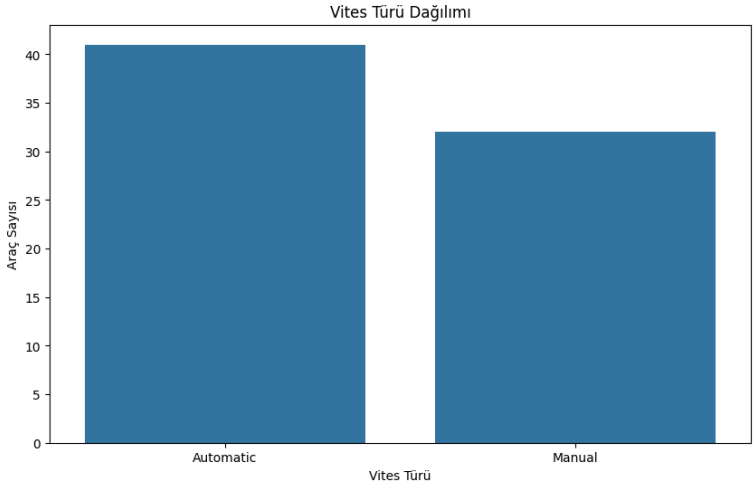
Veri setinde bulunan markalara göre araç sayılarını incelersek Hyundai ve Volkswagen markasında en çok araç sayısını görüyoruz, Mahindra model araç sayısı da en az olduğunun analizini yapabiliriz.



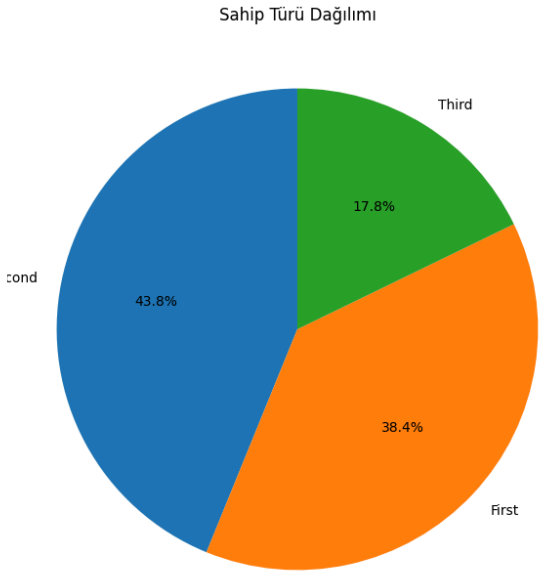
Yıllara göre veri setine eklenen araç sayını karşılaştırdık ve en çok araç sayısını 2019 değerinde 20+ olarak analiz yapabiliriz en az araç sayısı 2016 yıllarında kaydedilmiş.



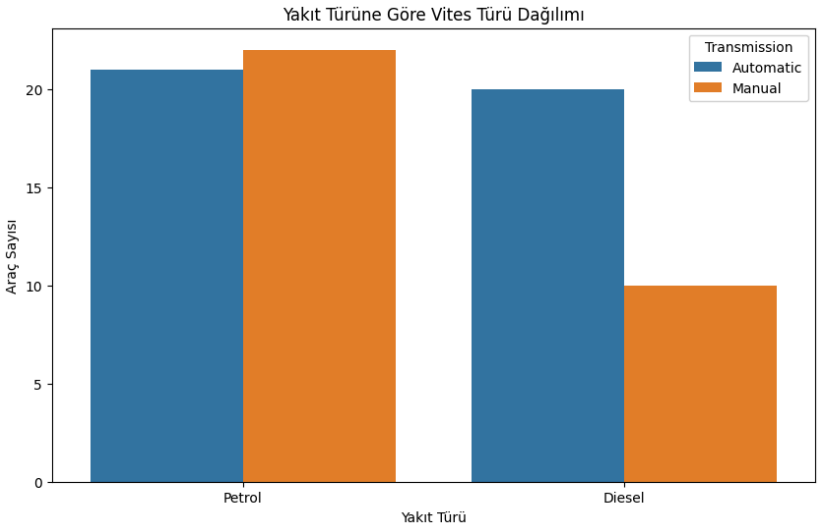
Veri setimizde bulunan araç sayısına göre yakıt türlerini karşılaştırdık ve sonuç olarak neredeyse yarı yarıya bir sonuç aldığımızı görebiliriz.



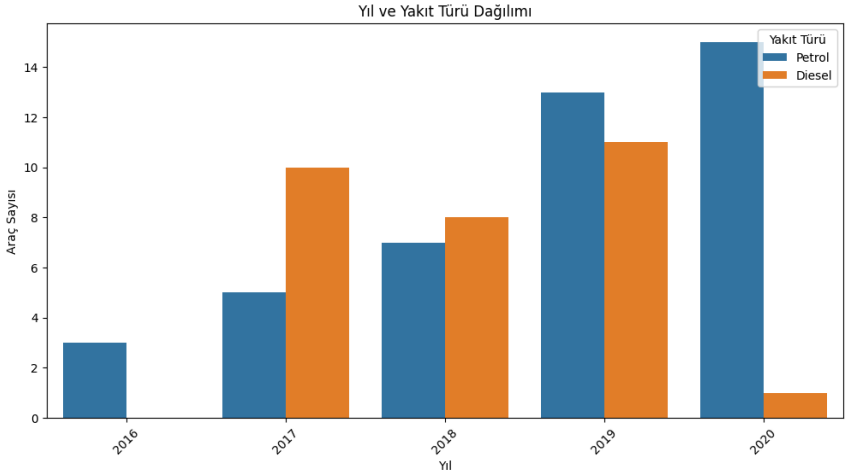
Araç sayısına göre vites türünün dağılımını inceledik ve sonuç olarak otomatik vitesi olan araçların daha fazla olduğunu analiz edebiliriz.



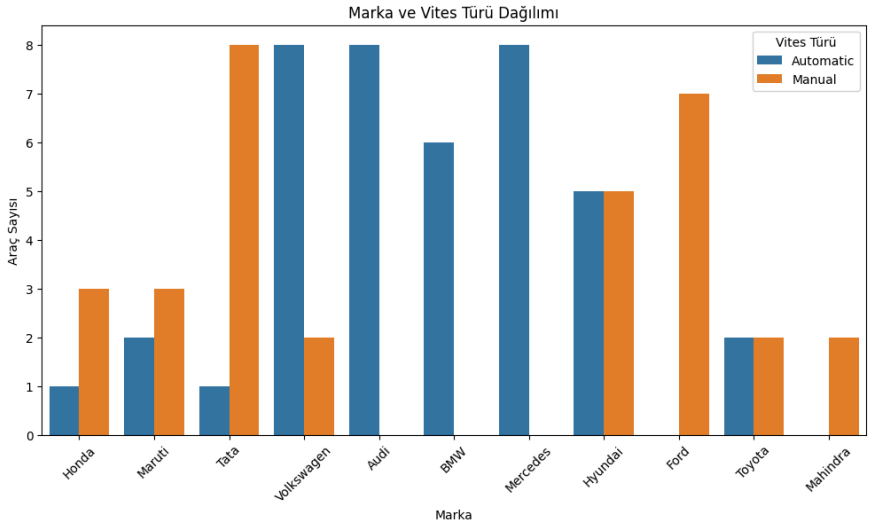
Veri setinde bulunan araçların kaçıncı el olduğunun pasta grafiğinde inceledik ve genel olarak 2. el araç sayısının daha fazla olduğunun analizini yapabiliriz.



Yakıt türüne göre vites türünü karşılaştırdık ve yakıt türü petrol olan araçların düz vites olduğunu, yakıt türü dizel olan araçların otomatik vites olduğunun analizini yapabiliriz.

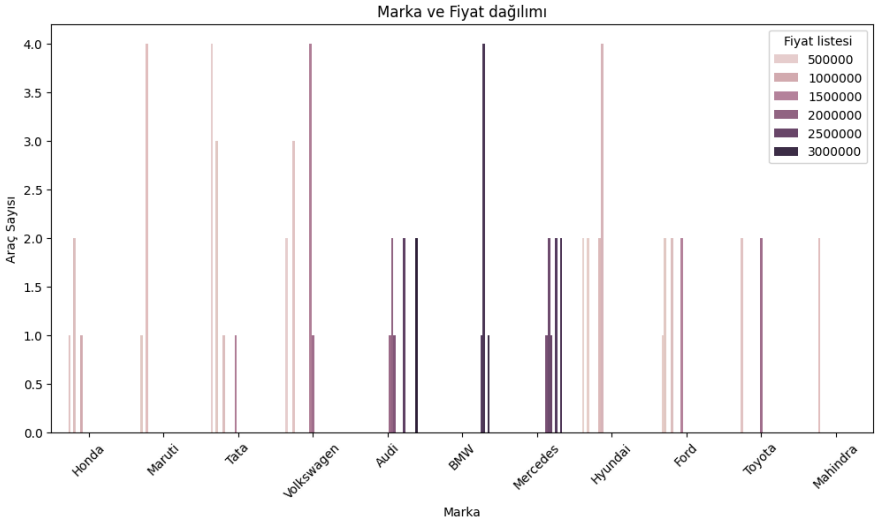


Yıla göre yakıt türünü karşılaştırdık ve 2016 yılında hiç dizel araç bulunmamakta ve 2017 yılında benzin kullanan araç sayısına göre dizel araçlar daha artışta fakat sonrasında bir azalım, son bir yükseliş sonrasında tamamen benzinli araçlara geri dönüldüğü analizi yapılabilir.



Bu veri setinde 2016 ve 2020 yılları arasında Audi, BMW, Mercedes markasında hiç vites türü manuel olan araç bulunmamakta ve Ford, Mahindra markasında vites türü otomatik olan araç bulunmamakta.

Tata markasında vites türü manuel olan araçların en fazla olduğunu görüyoruz ve otomatik vites türünde ki araç sayısı da en fazla Volkswagen, Audi, Mercedes markalarında olduğunun analizini yapabiliriz.



Bu analizimizde fiyat ve markaları karşılaştırdık ve en pahalı araçların BMW modelinde olduğunu, en ucuz araçların Mahindra, Maruti modellerinde olduğunun analizini yapabiliriz.

Veri seti Model Analizi

Ortalama Mutlak Hata : 276735.84383043397

Ortalama Kare Hata : 110547144655.49069

Kök Ortalama Kare Hata : 332486.3074706847

Tahmin Sonuçları

GERÇEK - TAHMİN

2200000 - 2367836.0
550000 - 386853.4
3200000 - 2981689.0
1000000 - 1503816.0
1500000 - 1828165.0
600000 - 586712.1
1800000 - 1791424.0
1800000 - 1571984.0
3000000 - 2391514.0
650000 - 889171.3
700000 - 1054986.0
700000 - 990339.5
650000 - 623728.7
1600000 - 2138442.0
2700000 - 2237813.0

SON OLARAK MODELİMİZİN İYİ BİR MODEL OLDUĞUNU AKTARABİLİRİZ ÇÜNKÜ FİYAT DEĞİŞKENİNİ TAHMİN ETMEYE ÇALIŞTIĞIMIZ İÇİN BAŞLANGIÇTA MODELİN HATA DEĞERLERİ ÇOK YÜKSEK GİBİ GÖZÜKEBİLİR FAKAT GAYET İYİ TAHMİNLERDE ÇOK UFAK SAPMALARDA BULUNUYOR.