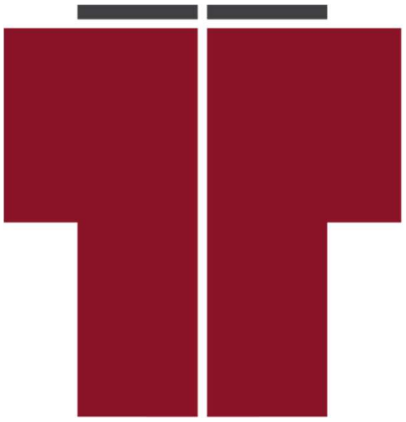


İSTANBUL
TOPKAPI
ÜNİVERSİTESİ

2025-2026 Güz Dönemi

Bölüm: Yazılım Mühendisliği

Ders: Veri Madenciliği (FET445)

Proje: Car Price Prediction Project

Grup: Motor Çetesi

Danışman: Yıldız Karadayı

Alper Ceylan – 22040301071

Ahmet Buğra Kaplan – 22040301028

Emirhan Sertel – 22040301064

Yusuf Utku Öztürk – 22040301013

GitHub: https://github.com/AhmetBugraKaplan/CarRecommendationSystem

25/11/2025

1)Problem Tanımı

1.1) İş / Bilimsel Soru: Kişiler araç alırken, satarken veya piyasa araştırması yaparken karşılaştığı fiyatların gerçeği yansıtıp yansıtmadığını bilememektedir. Çoğu zaman aynı özelliklere sahip araçlar arasında fiyat farkları olabilmekte ve bu fiyat farkları kişilerin karar verme süreçlerini zorlaştırmaktadır. Bu proje Craigslist.org sitesinden alınan 2.el araçların bilgilerini tutmaktadır. Böylece kullanıcıya veri temelli bir değerlendirme imkânı ve doğru fiyat aralığı sunmayı hedeflemektedir.

1.2) Görev Türü: Araba özelliklerini kullanarak fiyat tahmini yapan regresyon modeli.

1.3) Hedef Değişkenler: Projede kullanılan hedef değişken price olup araçların satış fiyatını temsil eder. Sayısal bir değişken olan bu değer USD cinsindedir ve modelin tahmin etmeye çalıştığı temel çıktıdır. Araçların sahip olduğu kategorik ve sayısal özelliklerden yola çıkarak fiyatın öngörülmesi amaçlanmaktadır. Veri setindeki fiyat dağılımı sağa çarpık olduğu için, regresyon modellerinin daha stabil ve etkili çalışması adına log1p(price) dönüşümü tercih edilmiştir. Proje bir regresyon problemi olduğundan, yalnızca tek bir hedef değişkeni bulunmaktadır ve herhangi bir pozitif ya da kategorik sınıf yapısı içermez.
1.4) Başarı Kriterleri: Araç fiyat tahmin modellerinin hedeflenen performansa ulaşabilmesi için belirli nicel ölçütler tanımlanmıştır. Modelin üreteceği tahminlerle gerçek fiyatlar arasındaki farkın ortalama mutlak değeri 3000 doların altında olmalıdır. Aynı zamanda uç değerlerin modele etkisini gösteren RMSE değerinin de 4500 doların altında kalması beklenmektedir. Buna ek olarak, modelin toplam varyansın en az yüzde 80'ini açıklaması gerektiği için R^2 değerinin 0.80'in altına düşmesi düşük performans olarak kabul edilmektedir. Bu üç metrik birlikte değerlendirildiğinde, modelin hem ortalama hem de uç noktadaki hatalarının kontrol altında olması ve genel tahmin gücünün yüksek olması amaçlanmaktadır.

2) Proje Yönetimi

2.1) Kilometre Taşları ve Zaman Çizelgesi:

1. Hafta: Veri seti ve proje konusu seçimi: Projenin ilk haftasında ekip oluşturulmuş ve proje konusu "Araç Fiyat Tahmini + Araç Öneri Sistemi" olarak belirlenmiştir. Bu dönemde Craigslist araç veri seti, proje için temel veri kaynağı olarak seçilmiştir.(Sorumlu kişiler: Tüm ekip)

2. Hafta: Veri Ön İşleme, Veri Hazırlama ve EDA Hazırlıkları: İkinci hafta boyunca eksik değer analizi yapılmış, tekrarlanan kayıtlar temizlenmiş ve price ile odometer değişkenlerindeki aykırı değerler incelenmiştir. Bu aşama veri kalitesini ölçmek ve EDA'ya temel oluşturmak amacıyla yürütülmüştür.(Sorumlu kişiler: Alper Ceylan ve Ahmet Buğra Kaplan)

3. Hafta: Keşifsel Veri Analizi (EDA) ve Görselleştirmeler: Üçüncü haftada kapsamlı EDA gerçekleştirilmiştir. Price, odometer ve year değişkenlerinin dağılımları analiz edilmiş; korelasyon matrisi ve Mutual Information (MI) sonuçları değerlendirilmiştir. Kategorik değişkenlerin frekans dağılımları oluşturulmuş ve veri setinin genel yapısı detaylandırılmıştır.(Sorumlu kişiler: Emirhan Sertel ve Yusuf Utku Öztürk)

4–5. Haftalar: Model Geliştirme ve Hyperparameter Tuning: Dördüncü haftada preprocessing pipeline geliştirilmiş; one-hot encoding, label encoding, scaling işlemleri ve model pipeline entegrasyonu uygulanmıştır.

Beşinci haftada Linear Regression, Decision Tree ve SVR modelleri eğitilmiş; GridSearchCV ile hiperparametre aramaları yapılmış ve modeller MAE, RMSE ve R^2 metrikleri üzerinden karşılaştırılmıştır. (Sorumlu kişiler (Pipeline + EDA Destekleri): Emirhan Sertel ve Yusuf Utku Öztürk + Sorumlu kişiler (Modelleme + Tuning): Alper Ceylan ve Ahmet Buğra Kaplan)

6. Hafta: Performans Analizi ve Değerlendirmesi: Bu aşamada seçilen modellerin performansı detaylı şekilde değerlendirilmiş, hiperparametre tuning sonuçları analiz edilmiş ve en iyi model belirlenmiştir. Model metrikleri (MAE, RMSE, R^2) karşılaştırılmış ve nihai sonuçlar doğrulanmıştır.(Sorumlu kişiler: Tüm ekip)

7. Hafta: Nihai Raporun Hazırlanması ve Sunumun Oluşturulması: Son hafta proje raporu son hâline getirilmiş, slayt sunumu hazırlanmış ve ekip içi görevlerin doğrulanması sağlanmıştır. Tüm analizler, grafikler, pipeline yapıları ve model kıyaslamaları sunuma entegre edilmiştir. (Sorumlu kişiler: Tüm ekip)

2.2) Roller ve Sorumluluklar (Şablona Uygun Özet)

Ahmet Buğra Kaplan: Projenin teknik yükünü üstlenmiştir. Preprocessing pipeline tasarımı, encoding ve scaling stratejilerinin belirlenmesi, model pipeline entegrasyonu, Linear Regression–Decision Tree–SVR modellerinin kurulumu, GridSearchCV/RandomizedSearchCV ile hiperparametre optimizasyonu ve performans karşılaştırmalarından sorumludur. GitHub mimarisinin yönetimini de gerçekleştirmiştir.
Alper Ceylan: Proje görev dağılımında veri temizleme, eksik değerler ve baseline modeller kısmı kendisine bırakılmıştır. Veri seti içerisindeki eksik veri analizi ve uygun imputasyon stratejilerinin uygulanması, tutarsız ve aykırı değerlerin tespit edilmesi, veri temizliği adımlarının belgelenmesi görevlerini üstlenmiştir. Baseline modellerin (Dumb baseline ve basit regresyon modelleri) hazırlanmasından sorumlu olmuş ve model sonuçlarının ilk değerlendirmelerine katkı sağlamıştır.
Emirhan Sertel: EDA'nın ana sorumlularındandır. Dağılım grafikleri (price, odometer, year), korelasyon ve MI analizleri, kategorik değişken görselleştirmeleri, aykırı değer analizleri ve EDA rapor bölümünün yazımı kendisine aittir. Sunumdaki birçok grafik onun tarafından hazırlanmıştır.
Yusuf Utku Öztürk: Destekleyici analitik görevleri üstlenmiştir. Değişkenlerin derinlemesine incelenmesi, özellik mühendisliği önerileri, model sonuçlarının doğrulanması, GridSearchCV çıktılarının tablo/grafik hâline getirilmesi ve sunumun performans kısmının hazırlanması görevlerini yürütmüştür.

Zaman çizelgesine uyum, kod bütünlüğünün korunması, nihai rapor kontrolü ve sunum hazırlıkları ekip tarafından birlikte yürütülmüştür.

2.3) Çıktılar:

Proje sonunda oluşacak çıktılar ve bunların GitHub linkinde bulunduğu konumlar:

1. Final Proje Raporu (PDF): Projenin tüm aşamalarını detaylı olarak açıklayan kapsamlı rapor

2. Jupyter Notebook Kod Dosyaları: Her grup üyesinin geliştirdiği modelleri içeren notebook'lar

    22040301071.ipynb (Alper Ceylan)

    22040301028.ipynb (Ahmet Buğra Kaplan)

    22040301064.ipynb (Emirhan Sertel)

    22040301013.ipynb (Yusuf Utku Öztürk)

3. Veri Seti ve Veri Sözlüğü: Craigslist Cars & Trucks Dataset

5. GitHub Repository: https://github.com/AhmetBugraKaplan/CarRecommendationSystem

3) İlgili Çalışmalar (Mini Literatür incelemesi): Proje geliştirme aşamasında aynı veri seti ile yapılan benzer projelerden farklı bakış açısı kazanılması amacıyla grup olarak incelemelerde bulunuldu. İncelenen projelerde kullanılan modeller ve teknikler bizim gerekliliklerimiz açısından analiz edildi ve kendi projemiz ile karşılaştırmaları yapıldı. İncelenen projeler şu şekildedir:

Used Car Price Prediction With Tree Based Models (Kaggle): Ahmed (2023) Kaggle çalışmasında bizimle aynı Craigslist veri setini (426,000+ kayıt) kullanarak XGBoost, LightGBM ve CatBoost modelleri ile stacking tekniği uygulamıştır [3]. Optuna ile hiperparametre optimizasyonu yapılmış ve 0.95-0.98 arası R^2 skorları elde edilmiştir.

Araç Fiyat Tahmininde Makine Öğrenmesi Algoritmalarının Karşılaştırılması Ve Performans Analizi (DergiPark Akademik): Kıvrak (2025) çalışmasında Türkiye otomotiv sektöründe 23,900 araç kaydı üzerinde 10 farklı makine öğrenmesi algoritması test etmiştir [1]. Lasso Regression %99 R^2 değeri ile en başarılı model olurken, Random Forest ve Gradient Boosting %97 R^2 ile güçlü alternatifler olmuştur. Veri seti Ocak 2022 - Aralık 2023 dönemini kapsamaktadır

Linear Regression- Car Price Prediction and Data Analysis (Medium): Hoxhaj Medium blog yazısında temel linear regression yaklaşımını açıklamıştır [2]. Yaklaşık 10,000 kayıtlık Kaggle veri setinde Linear Regression, Polynomial Regression ve Ridge Regression modelleri kullanılmış ve 0.75-0.80 arası R^2 skorları elde edilmiştir.

İncelenen çalışmalar farklı veri setleri kullanmıştır: Kıvrak 23,900, Hoxhaj 10,000, Ahmed 426,000 kayıt. Bizim projemiz 360,398 kayıt üzerinde çalışmıştır. Model çeşitliliği açısından Kıvrak 10 model, Hoxhaj 3 model, Ahmed 3 tree-based model kullanırken, bizim projemiz 8 farklı model ailesini test etmiştir.

Bizim projemizin temel farklılıkları şunlardır: Birincisi, condition değişkeninin kilometre bilgisine göre düzeltilmesi hiçbir çalışmada görülmemiştir ve %67 satırı etkilemiştir. İkincisi, 17 yeni feature ile literatürün en kapsamlı feature engineering'i yapılmıştır. Üçüncüsü, 8 farklı model ailesi (linear, tree-based, distance-based, kernel-based, Bayesian) karşılaştırılarak en geniş model perspektifi sunulmuştur. Dördüncüsü, Pipeline yapısı kullanılarak deployment-ready bir sistem geliştirilmiştir. Son olarak, hem log-scale hem de gerçek dolar değerlerinde metrikler hesaplanarak model performansı daha anlaşılır hale getirilmiştir.

Ahmed'in çalışması %95-98 R^2 elde ederken, bizim projemiz %74 R^2 elde etmiştir. Bu fark, bizim projemizde agresif veri temizleme (%15 veri elendi) ve overfitting kontrolü (max_depth=5 gibi kısıtlamalar) yapılmasından kaynaklanmaktadır. Hedefimiz maksimum R^2 değil, gerçek dünyada güvenilir çalışan dengeli bir model oluşturmaktır.

4) Veri Tanımı ve Yönetimi:

4.1) Veri Seti:

Ad: Craigslist Cars & Trucks Dataset
Kaynak: Craigslist.org, web scraping ile toplanmış ABD'de kullanılmış araç ilanları.
Bağlantı: https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data
Lisans: CC0 Public Domain - Akademik ve araştırma amacıyla serbestçe kullanılabilir.

4.2) Veri Şeması:

426,880 satır × 26 sütun (~1.45 GB CSV)içermektedir.

Temel değişkenler: id (unique), price (USD, sürekli), year (1920-2020), manufacturer (43 kategori), model (string), condition (6 ordinal), cylinders (3-12), fuel (gas/diesel/electric/hybrid),

odometer (miles), title_status (clean/salvage vb.), transmission (automatic/manual), drive (fwd/rwd/4wd), type (sedan/SUV/pickup vb. 13 kategori), paint_color (12 renk), lat/long (konum), description (text), state/region (coğrafi).

Eksik değerler: size (%71), drive (%30), paint_color (%30), type (%22).

4.3) Boyut:426,880 ilan, 26 özellik. Dengesiz dağılım: Ford dominant (%15+), gas yakıt %90, fiyat $5K-30K arası yoğunlaşmış. Veri setinin ~%40-50'si eksik değer içeriyor.

4.4)Veri Erişim Planı: Verisetini kaggle linki üzerinden direkt bilgisayarımıza depoladık sonrasında bilgisayarımız üzerinden eriştik.

4.5)Etik, Gizlilik, Önyargı: Datasette herhangi bir kullanıcının kişisel verisi bulunmamaktadır bu sebeple KVKK'u ihlal eden herhangi bir durum bulunmamaktadır.

5) Keşifsel Veri Analizi (Exploratory Data Analysis)

5.1) Veri Kalitesi Kontrolleri: EDA aşamasında ilk olarak veri setindeki genel kalite sorunları incelenmiştir. Bu kapsamda tamamen aynı satırların tespiti için *df.duplicated()* fonksiyonu kullanılmış ve kaç adet tekrar bulunduğu belirlenmiştir. Aynı zamanda kullanılmayan sütunlar analiz edilerek gereksiz veri yoğunluğu oluşturan alanlar belirlenmiştir. Özellikle *id*, *image_url*, *description*, *posting_date*, *paint_color* ve *url* sütunlarının fiyatla ilişkili olmadığı, yüksek oranda eksik ya da model doğruluğunu olumsuz etkileyecek yapıda veriler içerdiği görülmüştür. Bu gözlemler veri hazırlama aşamasında temizlenecek sütunların belirlenmesini sağlamıştır.

Eksik değerlerin genel oranı ayrıca grafikler üzerinden incelenmiştir. Görseller, özellikle *cylinders*, *condition*, *drive*, *type* ve *fuel* sütunlarında dikkat çekici seviyede eksik değer bulunduğunu ortaya koymuştur. Bu tespitler, daha sonraki aşamalarda hangi sütunlara imputasyon uygulanması gerektiğini netleştirmiştir.

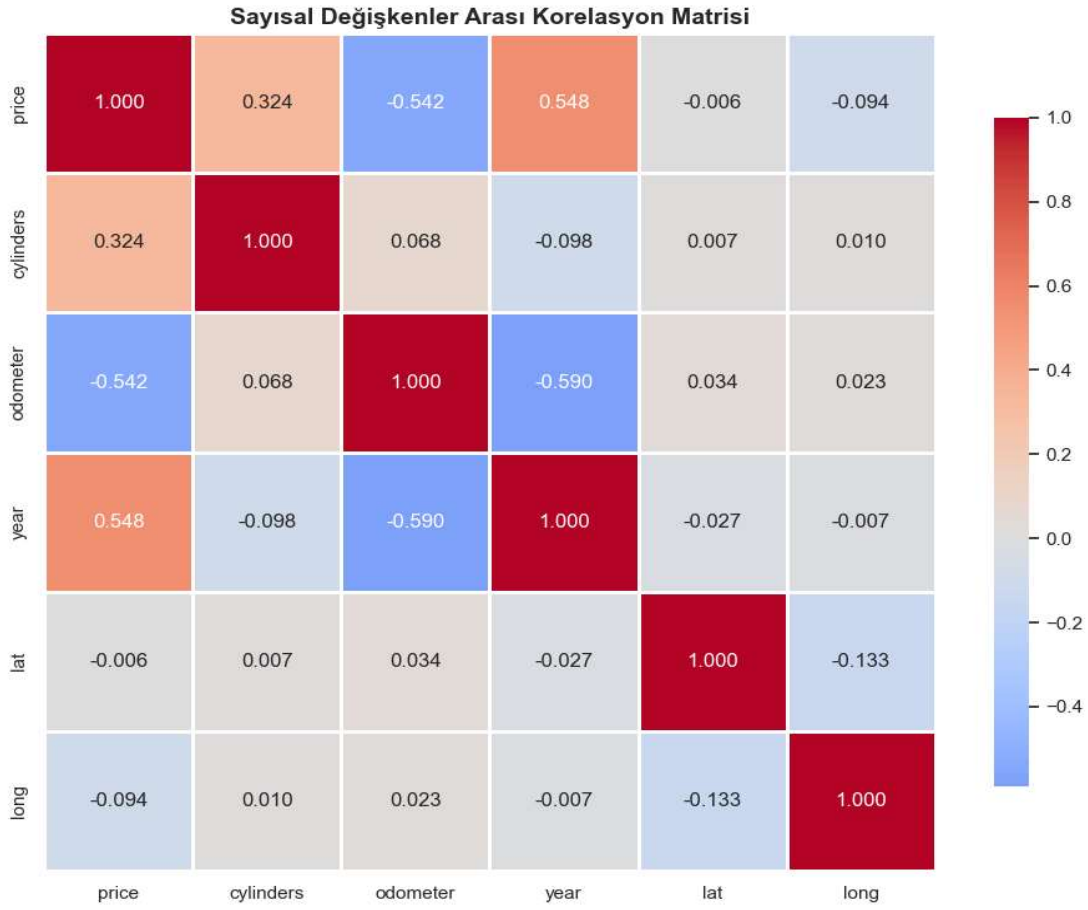
5.2) Dağılımlar ve Denge Analizleri: Hedef değişken olan *price* için histogram ve boxplot grafiklerinden yararlanılarak fiyat dağılımı incelenmiştir. Bu analiz, hem aykırı değerlerin tespit edilmesini hem de veri setinin sağa çarpık yapısının görülmesini sağlamıştır. Ardından *describe()* fonksiyonu ile fiyatın temel istatistiksel değerleri (min, max, mean, std, çeyreklikler) değerlendirilmiştir.

Fiyat dağılımında mantıksız değerlerin bulunması üzerine 500–500.000 dolar aralığının gerçekçi eşiği temsil ettiği anlaşılmış ve bu aralığın dışındaki gözlemlerin aykırı olduğu belirlenmiştir (bu temizlik işlemi veri hazırlama aşamasında uygulanmıştır).

Yıl değişkeninin dağılımı da hem scatter plot hem boxplot ile analiz edilmiştir. Grafiklerde özellikle gelecek yıllara ait tutarsız kayıtlar ve çok eski araçların fiyat davranışı net şekilde görülmüştür. Ayrıca 1980 altı araçların klasik/hurda kategorisine girdiği, fiyat dağılımını bozduğu ve model açısından problem yarattığı anlaşılmıştır (temizlik işlemi veri hazırlamada uygulanmıştır).

Odometer değişkeninin histogramı ve kilometre-fiyat scatter plot'u incelenerek yüksek kilometrelerin fiyat üzerindeki negatif eğilimi gözlemlenmiştir. Aykırı kilometre değerlerinin dağılımı ayrıca analiz edilerek hangi aralıkların model açısından problem yaratacağı belirlenmiştir.

Kategorik değişkenlerin dağılımları için barplot grafikleri hazırlanmış, ilk 10 kategori listelenmiş ve dengesizliklerin model öğrenmesini nasıl etkileyebileceği üzerine değerlendirme yapılmıştır.



5.3) Özellik – Hedef İlişkileri: Sayısal değişkenler arasındaki ilişkileri incelemek için korelasyon analizi gerçekleştirilmiştir. Price değişkeninin diğer değişkenlerle olan ilişkisi özellikle değerlendirilmiş, güçlü ve zayıf korelasyonlar belirlenmiştir. Ardından Mutual Information analizi ile tüm kategorik ve sayısal sütunların price ile ilişkisi ölçülmüş, değişkenlerin önem sıralaması çıkarılmıştır. En etkili ilk beş değişken ayrıca raporlanmıştır.

Pairplot görselleştirmesi ile sayısal değişkenlerin dağılımları, kümelenme yapıları ve olası aykırı noktalar detaylı olarak incelenmiştir.

5.4) Görselleştirme Planı: EDA boyunca çeşitli grafikler hazırlanmış ve sunum materyaline dahil edilmiştir. Bu grafikler;

- histogramlar (price, odometer, kategorik değişkenler),
- boxplotlar (price ve year aykırı değerleri),
- scatter plot ve trend çizgileri (year-price, odometer-price),
- pairplotlar,
- top-10 kategori barplotları

şeklindedir. Bu görseller EDA'nın temel analiz aşamalarını desteklemiştir.

6) Veri Hazırlama Planı

6.1) Temizleme Adımları: EDA'da belirlenen kalite sorunlarının ardından veri seti modellemeye uygun hâle getirilmiştir. Öncelikle tamamen aynı satırlar silinmiş ve tekrarlı kayıtlar ortadan kaldırılmıştır. Ardından gereksiz sütunların kaldırılması uygulanmıştır. *id, image_url, description, posting_date, paint_color, url* sütunları fiyat tahminiyle ilişkili olmadığı ve yüksek oranda eksik/yanıltıcı olduğu için çıkarılmıştır.

Yıl filtrelemesi kapsamında 1980 yılından önce üretilmiş araçlar veri setinden temizlenmiştir; bu araçların hem fiyat davranışı hem pazar yapısı modern araçlarla uyumsuz olduğundan modele zarar verdiği görülmüştür.

Fiyat değişkeninde 500–500.000 dolar aralığı dışında kalan değerler aykırı kabul edilerek çıkarılmıştır. Benzer şekilde odometer değişkeninde 300.000 mil üzeri araçlar %1'lik bir uç grup oluşturduğundan veri bütünlüğünü bozduğu için kaldırılmıştır. Bu adımların ardından *df_eda* üzerinden yeni bir temiz kopya oluşturulmuştur.

6.2) İmputasyon Stratejisi: Eksik değerlerin analizi sonrası uygun doldurma stratejileri belirlenmiştir. İlk olarak kategorik sütunlardaki anlamsız "other" değerleri *NaN* olarak dönüştürülmüş, böylece daha doğru bir şekilde imputasyon uygulanabilmiştir.

Ardından marka-model bazlı doldurma yapılarak *type*, *drive*, *condition*, *fuel* gibi sütunlardaki eksik alanlar ilgili kategorinin en sık görülen değeriyle tamamlanmıştır. Bu yöntem rastgele değil, veri yapısını koruyan mantıklı bir doldurma stratejisi olarak tercih edilmiştir.

Cylinders sütunundaki değerler "6 cylinders" gibi string formatta olduğundan sayısal formata dönüştürülmüş ve eksik değerler mod/median yaklaşımıyla tamamlanmıştır.

Year sütunundaki eksik veriler üretici (manufacturer) bazında medyan yıl değeri ile doldurulmuştur. Bu adımın ardından manufacturer sütunu artık gerekli olmadığından çıkarılmıştır.

Odometer sütunundaki eksik veriler için üretici–tip–yıl kombinasyonlarına göre gruplandırılmış medyan kilometre değeri kullanılmıştır. Doldurma sonrası veri sızıntısını önlemek amacıyla EDA ve modelleme için farklı veri kopyaları korunmuş ve EDA kopyasında doldurma sonrası odometer sütunu çıkarılmıştır.

Title_status sütunundaki eksik değerler ise mode yöntemiyle doldurulmuş, en sık görülen kategori referans alınmıştır.

6.3) Dönüşümler: Cylinders değerleri sayısal forma dönüştürülmüş, böylece modelin string verilerle çalışmasının önüne geçilmiş ve işlem verimliliği artırılmıştır.

6.4) Özellik Mühendisliği: Araç kondisyonu sütununun subjektif yapısı nedeniyle, kilometre bilgisiyle tutarlı olacak şekilde yeniden derecelendirme yapılmıştır. Yüksek kilometrede olup "excellent" görünen araçlar düzeltilmiş ve grafiklerle önce–sonra karşılaştırması yapılmıştır.

Odometer değerlerinin üretici–tip–yıl bazlı gruplanması sadece eksik veri doldurma değil aynı zamanda bir özellik mühendisliği adımı olarak değerlendirilmiştir.

Yıl değişkeniyle fiyat arasındaki ilişki tekrar incelenmiş, aykırı yılların çıkarılması sonrası grafiklerin daha tutarlı hâle geldiği doğrulanmıştır.

6.5) Özellik Seçimi ve Boyut İndirgeme: Korelasyon analizleri ve Mutual Information sonuçları birleştirilerek hangi değişkenlerin modele en fazla katkı sağladığı belirlenmiştir. Özellik önem sıralaması grafik ve tablolarla sunulmuş, güçlü/zayıf ilişkiler değerlendirilmiştir. Bu adımlar, model performansını artırmak için gerekli değişken seçiminin temelini oluşturmuştur.

7) Modelleme Planı:

7.1) Baseline Model Seçimi: Baseline model olarak projede bulunan her kişi farklı 2 model seçmiştir.Projede en yüksek puan almayı hedeflediğimiz ağaç tabanlı modellerde kişi bazlı yapılan değişikliklerin önemini görmek için herkes decision tree modelini kullanmıştır.

7.2) Model Ailesi seçimi:

Model Ailesi Seçimi Tablosu

Ekip Üyesi	Kullanılan Modeller	Kısa Gerekçe
Buğra	Linear Regression, Ridge, Decision Tree	Hızlı eğitim ve basit yapı; Decision Tree ile temel non-lineerlik testi.
Utku	KNN, Lasso (L1), Decision Tree	KNN ile komşuluk modelleme; Lasso ile feature selection; Decision Tree ile esneklik.
Alper	ElasticNet, Decision Tree	Dengeli regularization; Decision Tree ile non-lineer ilişkileri yakalama.
Emirhan	Linear Regression, KNN, Decision Tree	Baseline oluşturma; KNN ile lokal yapı; Decision Tree ile non-lineerlik testi.

8) Değerlendirme Tasarımı:

8.1) Kullanılan Metrikler: Proje bir regresyon problemi olduğu için model değerlendirmesinde MAE, MSE, RMSE, R^2 ve Adjusted R^2 metrikleri kullanılmıştır. Bu metrikler, hem ortalama hata davranışını hem de modelin açıklayabildiği varyans miktarını ölçerek performansı çok boyutlu olarak değerlendirme imkânı sunmuştur.

8.2) Doğrulama Protokolü: Model değerlendirme sürecinde veri, %80 eğitim ve %20 test olacak şekilde ayrılmıştır. Test seti yalnızca final performans ölçümünde kullanılmış ve modelin hiçbir aşamasında eğitime dahil edilmemiştir. Eğitim sürecinde Stratified 5-Fold Cross-Validation uygulanmış, böylece veri dengesizliği her katmanda korunarak modele yönelik olası bias azaltılmıştır.

Veri sızıntısını önlemek için tüm ön işleme adımları — scaling, encoding, imputasyon, feature selection ve hyperparameter tuning — yalnızca cross-validation döngüsü içinde çalıştırılmıştır. Tüm süreç sklearn pipeline yapısıyla kontrol altına alınmış, böylece test verisinin modele herhangi bir aşamada "öğretim verisi" olarak karışmasının önüne geçilmiştir.

Ayrıca random_state=42 sabitlenerek çalışmanın tekrarlanabilirliği güvence altına alınmıştır.

9) Riskleri Azaltma Yöntemleri

9.1) Veri Riskleri: Proje sürecinde veri setinde boyut dengesizliği, yüksek oranlarda eksik değerler ve kategorik değişkenlerde dağılım dengesizliği gibi sorunlarla karşılaşılmıştır. Bu problemler EDA aşamasında tespit edilmiş ve veri hazırlama sürecinde uygun yöntemlerle çözüme kavuşturulmuştur.

9.2) Azaltıcı Yöntemler: Model başarısız olma riskini azaltmak için çeşitli stratejiler uygulanmıştır. Öncelikle küçük prototip modellerle hızlı testler gerçekleştirilmiş ve başarılı yaklaşımlar geniş kapsamda uygulanmıştır. Feature importance analizleri ile model performansına katkısı düşük olan değişkenler elenmiştir. Model doğruluğunun yetersiz kaldığı senaryolarda ise alternatif yöntemler düşünülmüş; özellikle ensemble teknikleri (voting, stacking) gibi daha güçlü modeller potansiyel çözüm olarak değerlendirilmiştir.

10) Kullanılan Araçlar

10.1) Geliştirme Ortamı: Proje Python 25.3 sürümü üzerinde gerçekleştirilmiştir. Çalışmalarda numpy, pandas, matplotlib, seaborn ve sklearn kütüphaneleri yoğun şekilde kullanılmış; sklearn içerisindeki farklı modüller preprocess, modelleme ve değerlendirme süreçlerinde aktif rol oynamıştır. Reprodüksiyon için random seed değeri 42 olarak belirlenmiştir.

10.2) GitHub Deposu: Projenin tüm kaynak kodları ve dokümantasyonu aşağıdaki GitHub bağlantısı üzerinden erişilebilir durumdadır:

https://github.com/AhmetBugraKaplan/CarRecommendationSystem

11) Beklenen Sonuçlar ve Görselleştirme Planı

EDA aşamasından başlayarak özellik mühendisliği, model karşılaştırması ve performans değerlendirmesi bölümlerinde çok sayıda grafik üretilmiştir. Bu görseller, model davranışının hem veri düzeyinde hem de metrik düzeyinde etkili bir şekilde anlaşılmasına katkı sağlamaktadır. Tüm grafikler proje raporunda ilgili görsel bölümünde detaylı şekilde incelenebilir.
Proje boyunca oluşturulan tüm grafikleri "**14) Ek Grafik**" başlığı altında görebilirsiniz.

12)Referanslar:

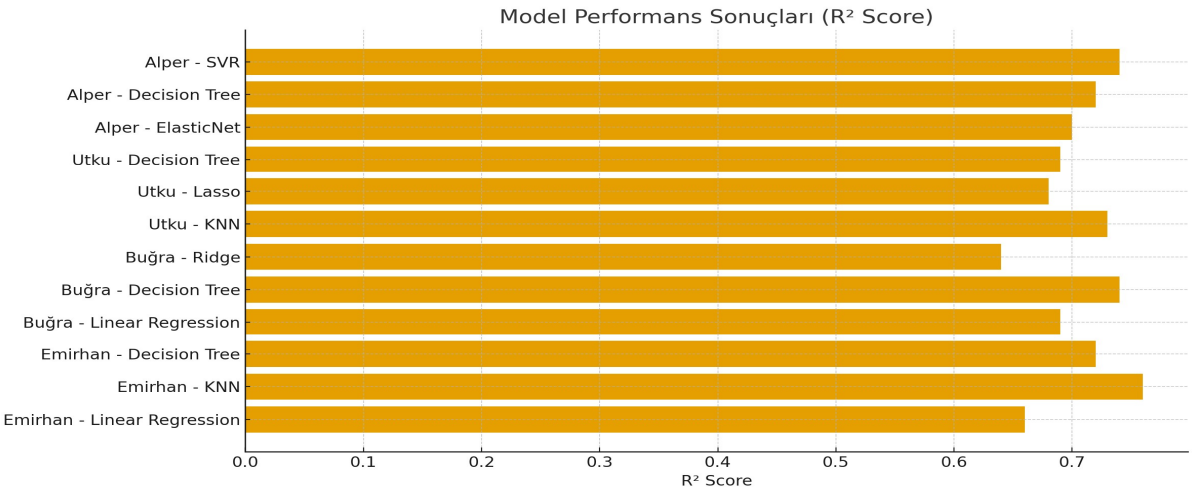
[1] O. Kıvrak, “Araç Fiyat Tahmininde Makine Öğrenmesi Algoritmalarının Karşılaştırılması ve Performans Analizi”, İKTİSAD, c. 10, sy. 27, ss. 454–474, 2025, doi: 10.25204/iktisad.1494020.

[2] D. Hoxhaj, "Linear Regression — Car Price Prediction and Data Analysis," Medium, 2023. [Online]. Available: https://medium.com/@diellorhoxhaj/linear-regression-car-price-prediction-and-data-analysis-112883cdd39b. [Accessed: Nov. 25, 2025].

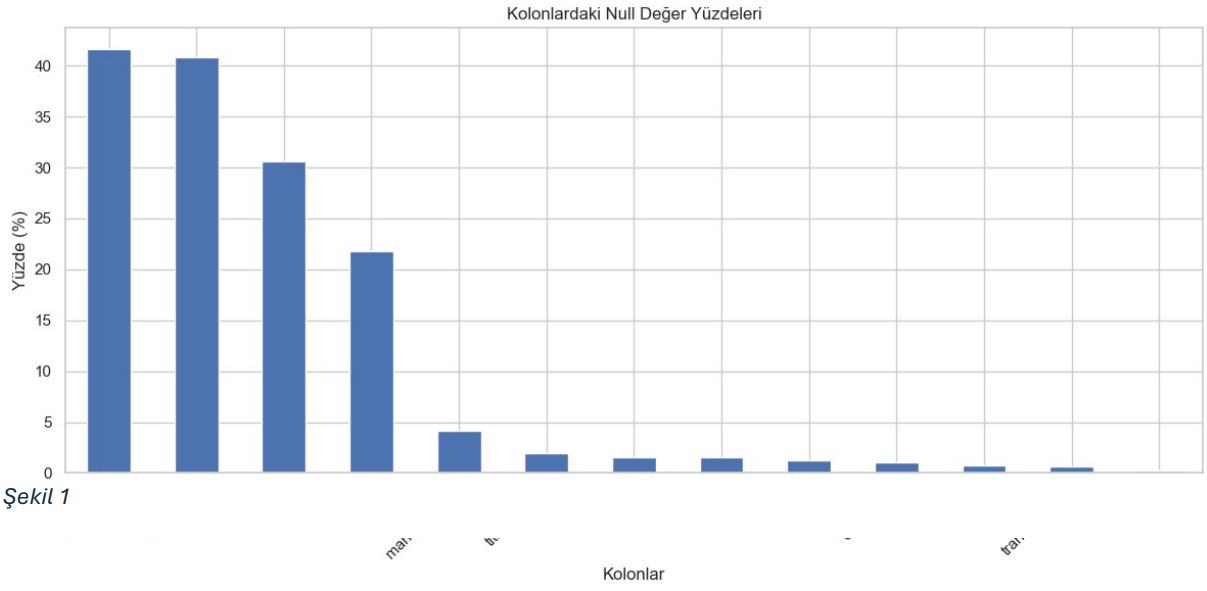
[3] A. Ahmed, "Used Car Price Prediction with Tree-Based Models & Stacking (Optuna Tuned)," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/code/abidahmed123/used-car-price-prediction-with-tree-based-models. [Accessed: Nov. 25, 2025].

[4] A. Reese, "Craigslist Cars & Trucks Dataset," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data. [Accessed: Nov. 25, 2025].

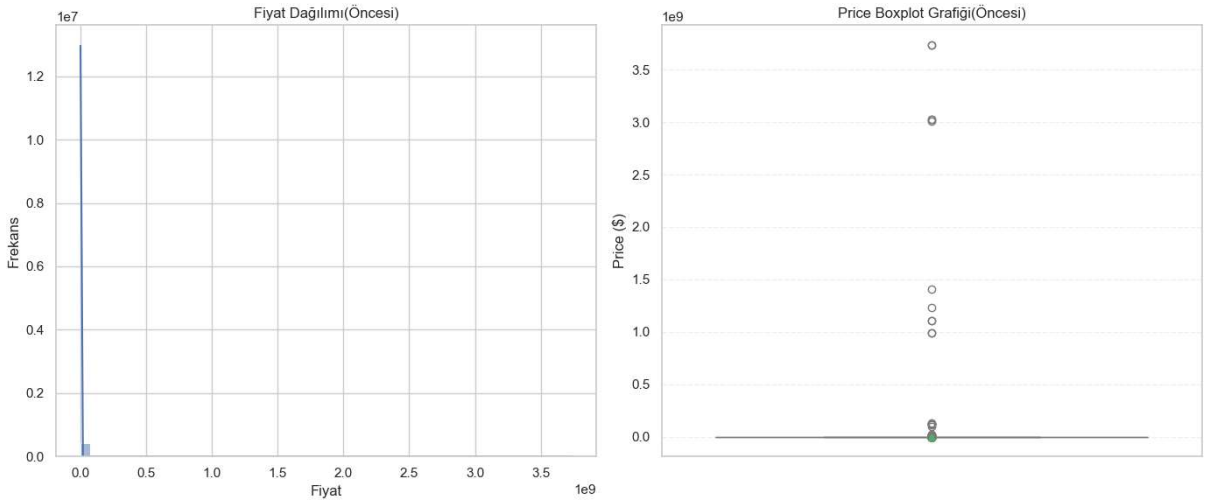
13)Bonus Kullanılan Tüm Baseline Modellerin R2 Sonuçları Karşılaştırma Tablosu



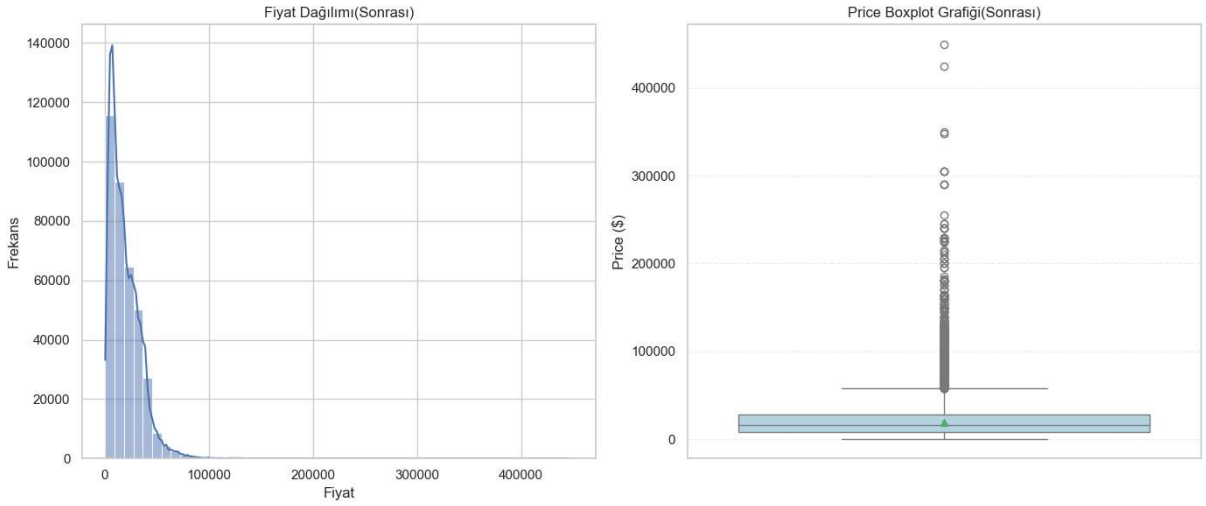
14) Ek Grafikler / Görseller



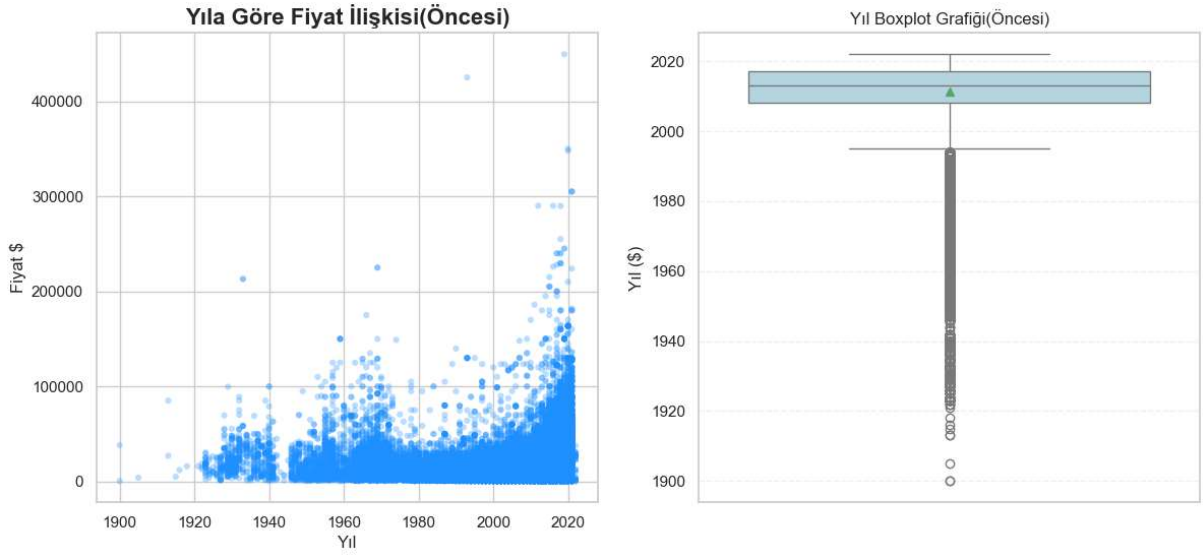
Şekil 1



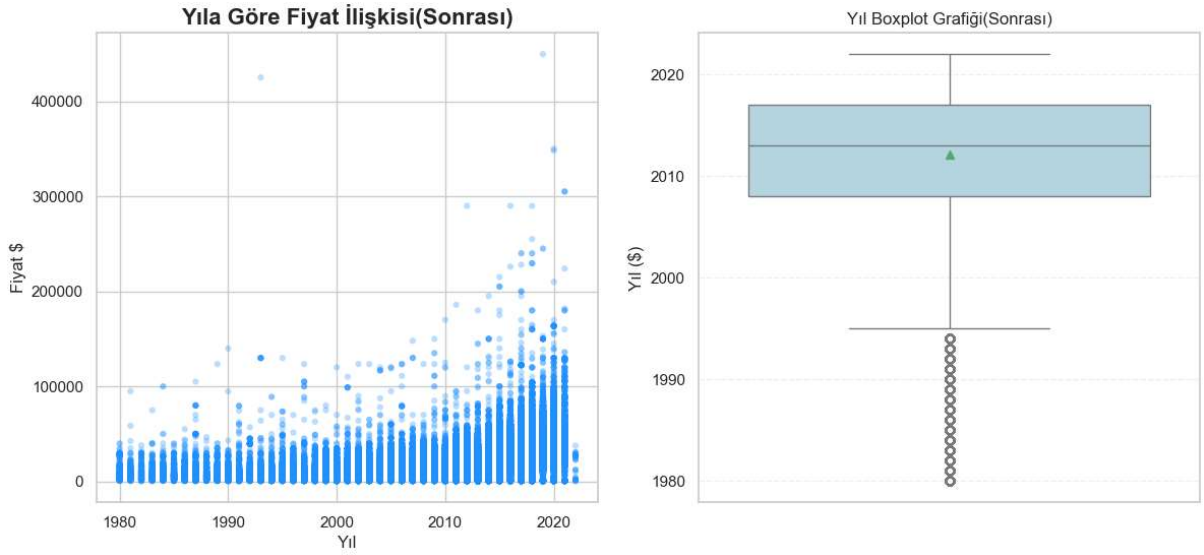
Şekil 2



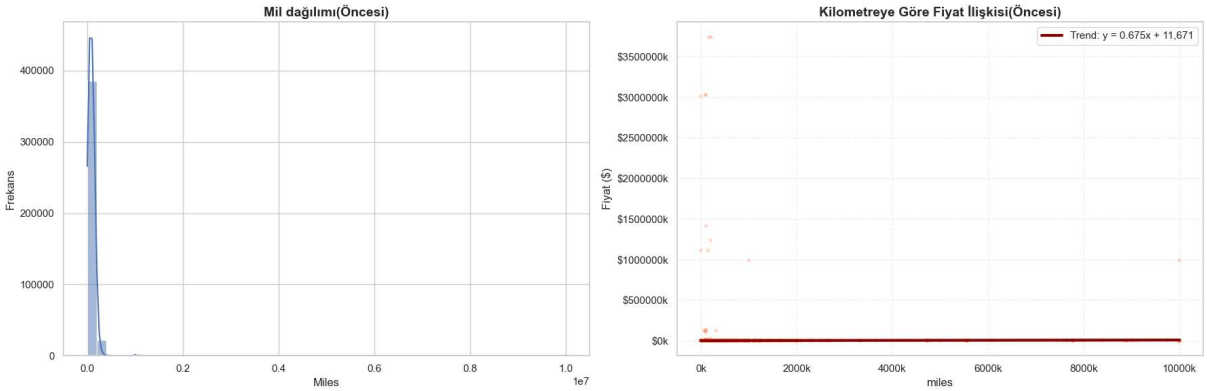
Şekil 3



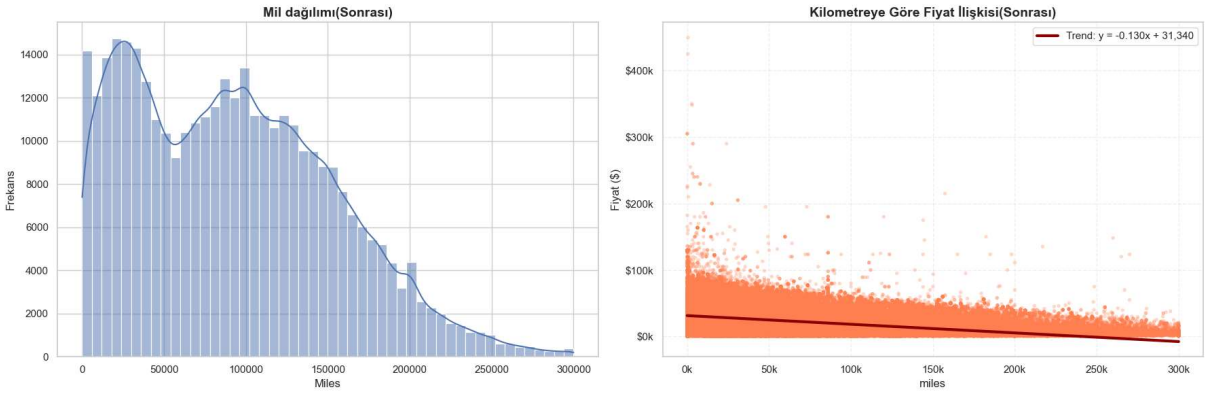
Şekil 4



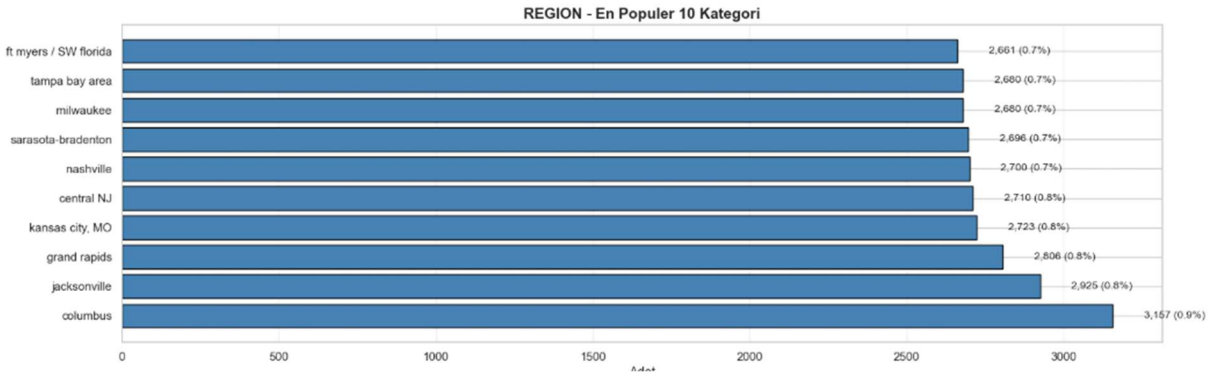
Şekil 5



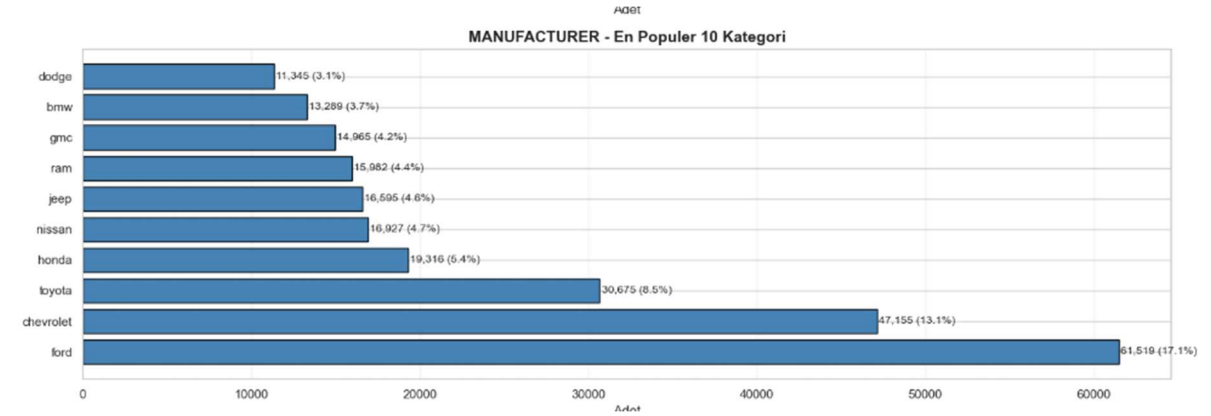
Şekil 6



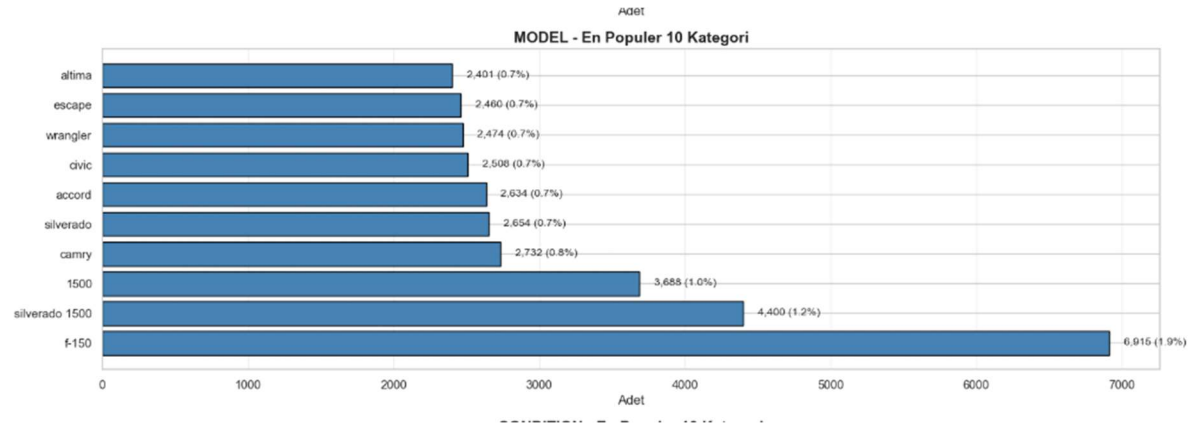
Şekil 7



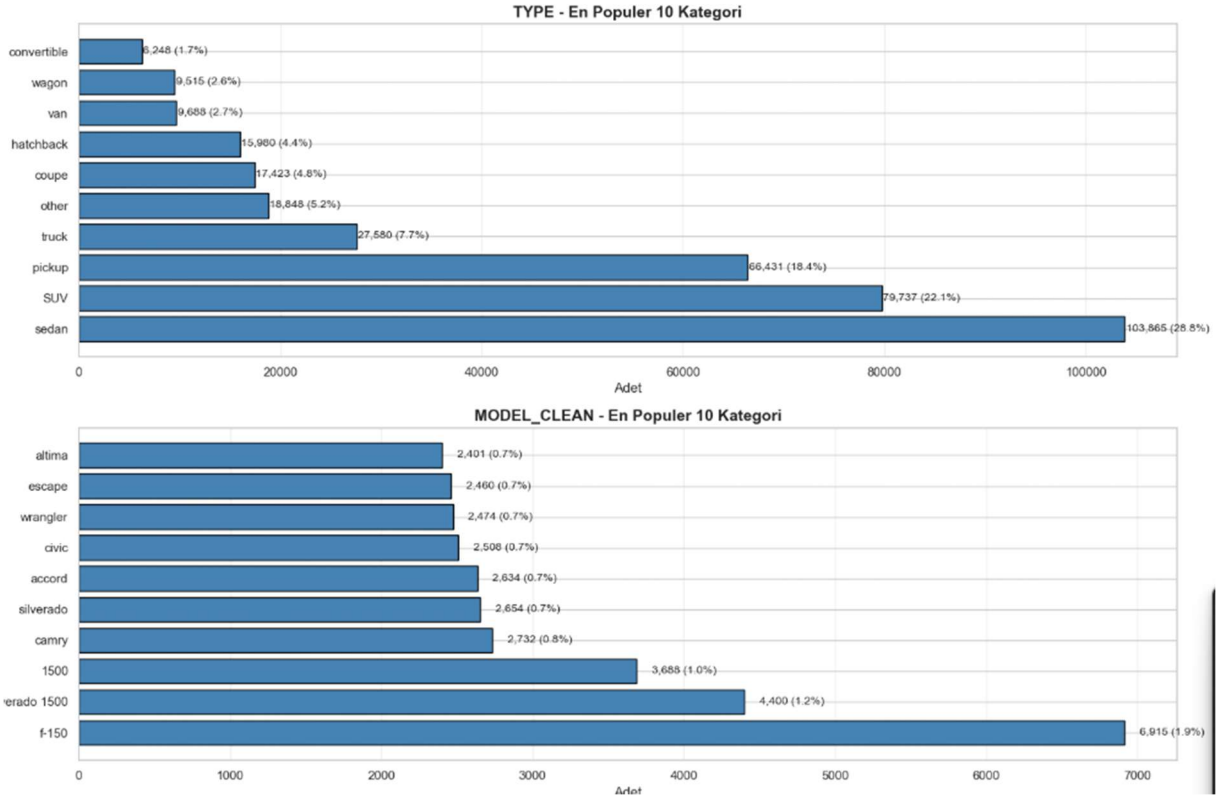
Şekil 8



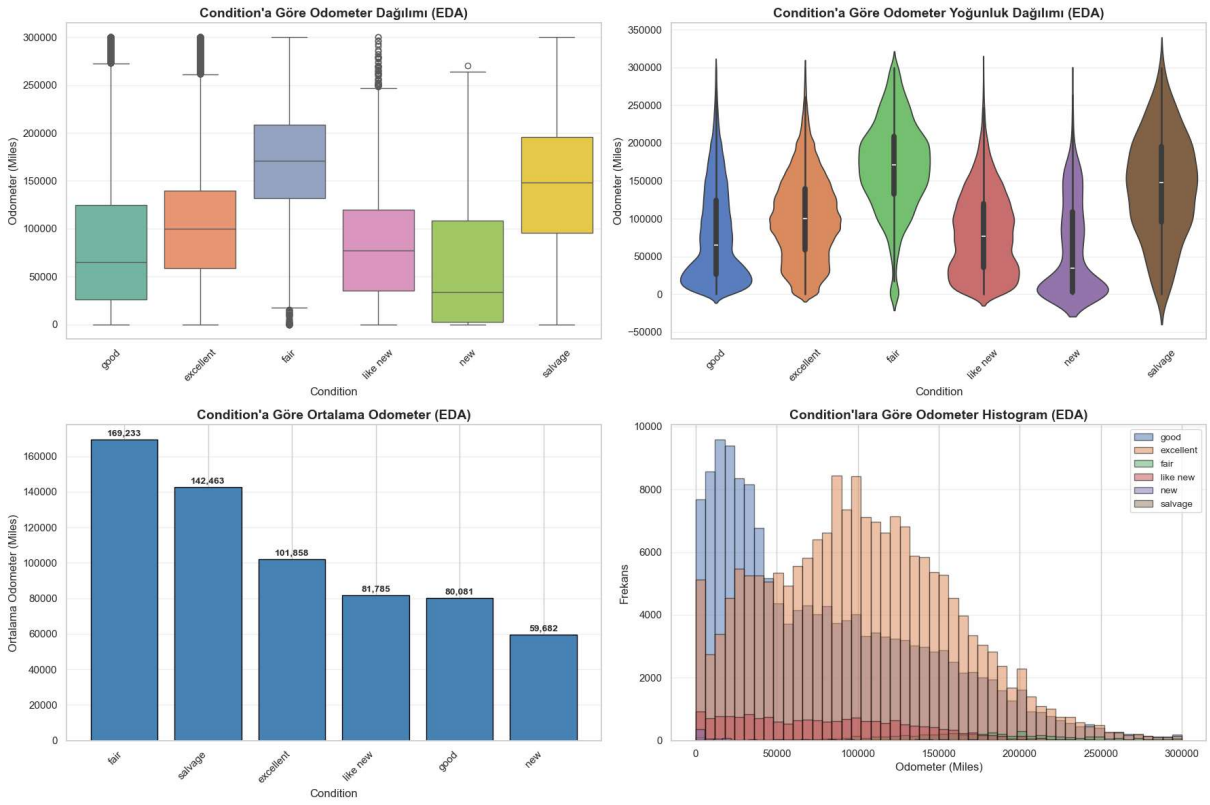
Şekil 9



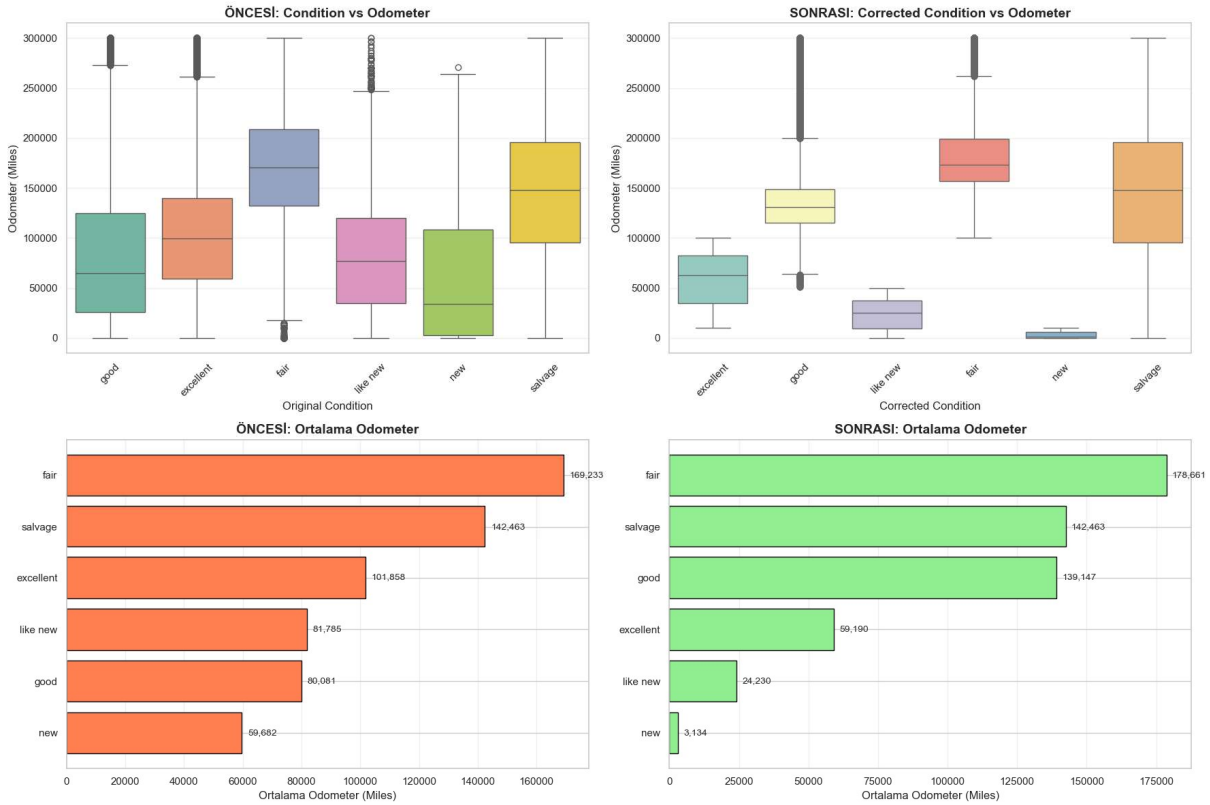
Şekil 10



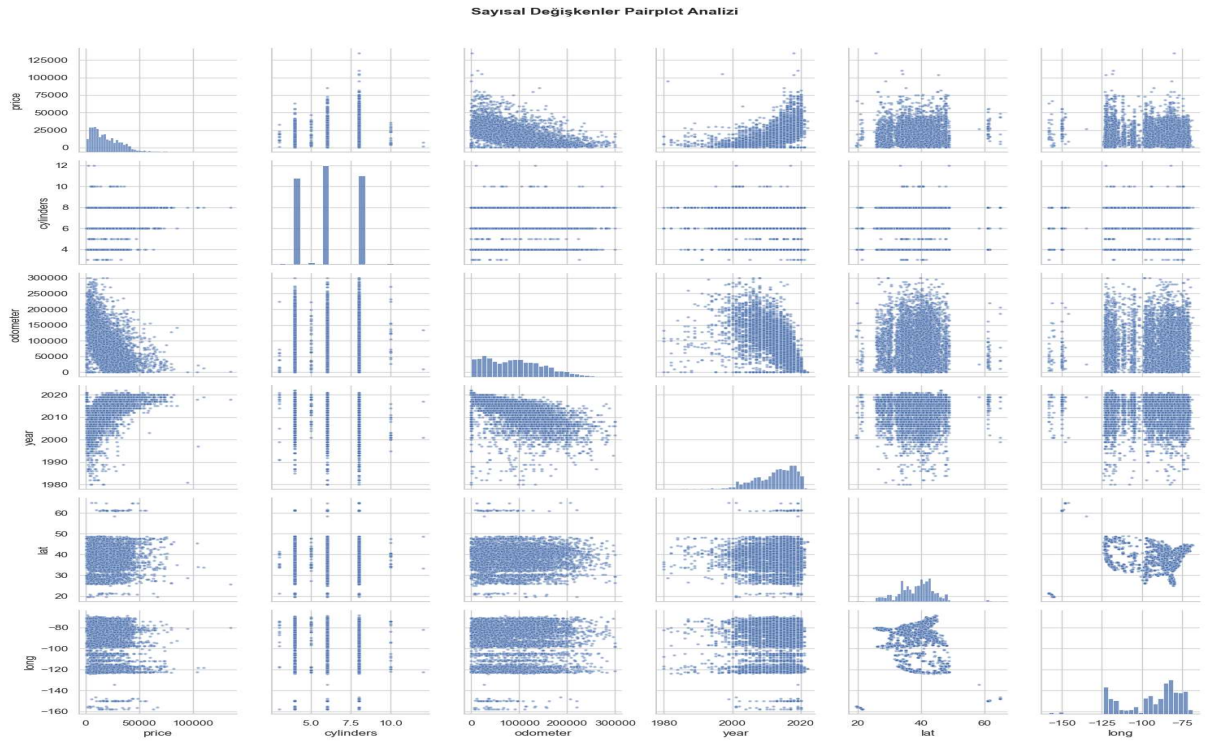
Şekil 11



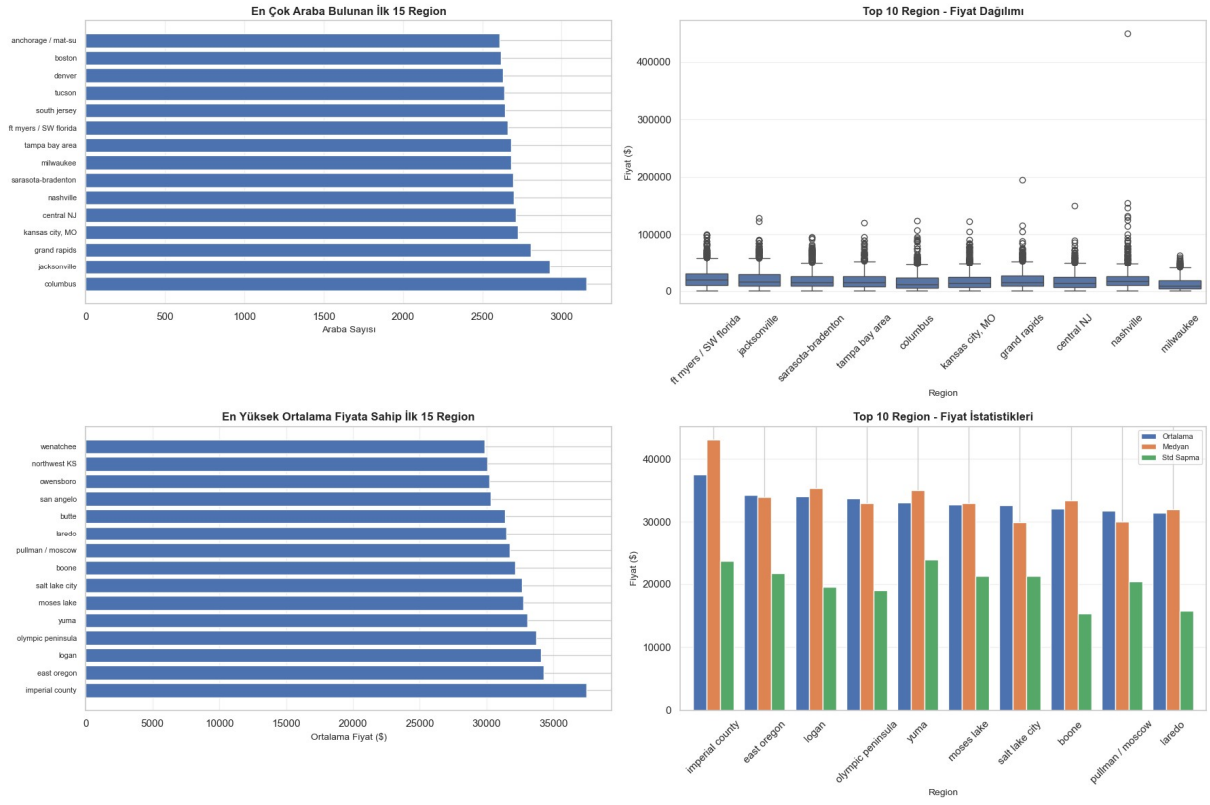
Şekil 12



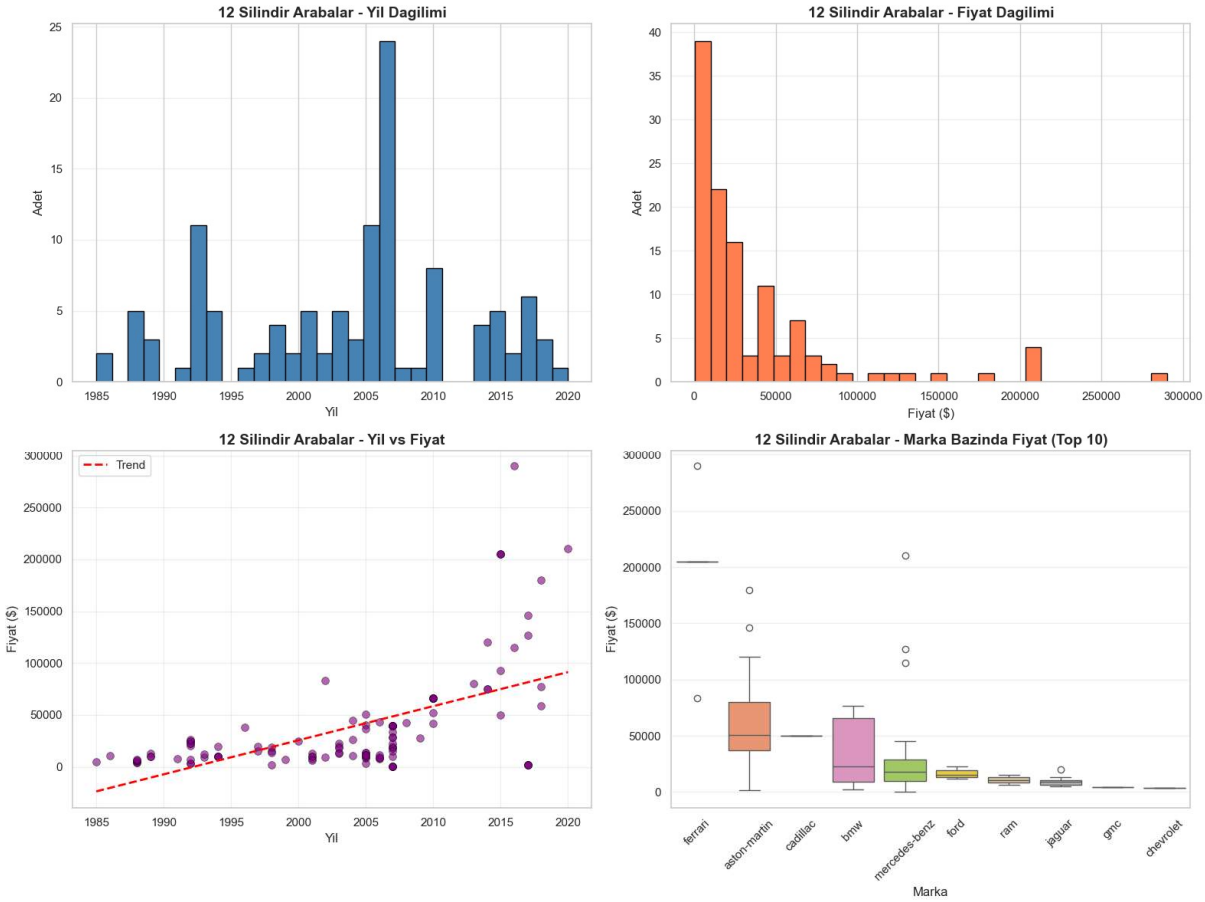
Şekil 13



Şekil 14



Şekil 15


12 Silindir Arabalar - Yil Dagilimi
Adet
Yil
12 Silindir Arabalar - Fiyat Dagilimi
Adet
Fiyat ($)
12 Silindir Arabalar - Yil vs Fiyat
Trend
Fiyat ($)
Yil
12 Silindir Arabalar - Marka Bazinda Fiyat (Top 10)
Fiyat ($)
ferrari
aston-martin
cadillac
bmw
mercedes-benz
ford
ram
jaguar
gmc
chevrolet
Marka

Şekil 16