Auto MPG Veri Seti ile Basit Doğrusal Regresyon Analizi Raporu

1.Proje Özeti ve Amaç

Bu proje raporu, Auto MPG veri setini kullanarak araç yakıt tüketimini tahmin etmek amacıyla gerçekleştirilen basit doğrusal regresyon analizini detaylandırmaktadır. Otomotiv sektöründe yakıt verimliliği, hem tüketici satın alma kararlarını hem de çevresel düzenlemeleri doğrudan etkileyen kritik bir parametredir. Bu bağlamda, araç tasarımı ve pazar analizine yönelik çalışmalarda, hızlı ve güvenilir tahmin araçlarına duyulan ihtiyaç her geçen gün artmaktadır. Bu çalışmanın temel hedefi, özellikle seçilen weight değişkeni ile mpg değişkeni arasındaki ilişkiyi modelleyerek, basit bir girdiden yola çıkıp pratik ve güvenilir bir tahmin aracı geliştirmektir. Proje kapsamında; veri ön işleme adımları, korelasyon analizi, temel regresyon modelinin oluşturulması, veri dönüşümü tekniklerinin uygulanması ve dönüştürülmüş modelin validasyonu gibi aşamalar titizlikle incelenmiştir. Özellikle, regresyon varsayımlarının sağlanması için uygun veri dönüşümlerinin belirlenmesi ve modelin tahmin performansının iyileştirilmesi üzerinde durulmuştur. Sonuç olarak, araçların temel bir özelliği olan weight değişkeni kullanılarak mpg değişkeninin başarılı bir şekilde tahmin edilmesi amaçlanmıştır.

2. Veri Seti Analizi

Bu çalışmada kullanılan Auto MPG veri seti, UCI Machine Learning Repository'den temin edilmiştir [1]. Veri seti, 1970'lerin sonları ve 1980'lerin başlarındaki otomobillerin yakıt tüketimi ve çeşitli özelliklerini içermektedir. Başlangıçta 398 gözlem ve 9 değişkenden oluşan veri seti, ön işleme adımları sonrasında 392 gözleme düşmüştür. Veri setindeki değişkenler ve ilk durumları aşağıdaki gibidir:

- mpg: mil/galon (bağımlı değişken)
- cylinders: silindir sayısı
- displacement: motor hacmi
- horsepower: beygir gücü
- weight: ağırlık
- acceleration: hızlanma
- model_year: model yılı
- origin: menşe ülke
- car_name: araba adı

Veri Ön İşleme Adımları:

1. *Gereksiz Değişkenlerin Çıkarılması:*
   car_name ve origin değişkenleri, regresyon analizi için doğrudan kullanılabilir nitelikte olmadıkları ve model karmaşıklığını artırabilecekleri için veri setinden çıkarılmıştır. car_name değişkeni benzersiz değerler içerirken, origin değişkeni ise her ne kadar uygun bir dönüşümle (örneğin one-hot encoding) modele dahil edilebilecek bir kategorik değişken olsa da, bu çalışmanın odak noktası yalnızca tek bir sayısal değişken üzerinden yakıt tüketimi

tahmini yapmak olduğundan kapsam dışında bırakılmıştır. Bu tercih, modelin yorumlanabilirliğini artırmak ve basit doğrusal regresyonun temel yapısına sadık kalmak amacıyla yapılmıştır.

2. *Eksik Veri Temizliği:*
   Veri setinde yalnızca horsepower değişkeninde eksik (NaN) değerler tespit edilmiştir. Bu eksik değerlerin sayısı toplamda 6 gözlem ile sınırlı olduğundan, ilgili satırlar veri setinden çıkarılmıştır. Temizlik sonrası veri seti 398 gözlemden 392 gözleme düşmüştür.

3. *Eğitim ve Test Ayrımı:*
   Temizlenmiş veri seti, modelin genellenebilirliğini değerlendirmek amacıyla %80 eğitim ve %20 test olmak üzere ikiye ayrılmıştır. Bu ayrım, random_state=123 parametresi kullanılarak tekrarlanabilirliği sağlanmıştır. Eğitim seti 313 gözlemden, test seti ise 79 gözlemden oluşmaktadır.

Veri setinin genel istatistiksel özetleri ve değişken tipleri, analiz öncesinde incelenerek veri yapısı hakkında bilgi edinilmiştir.

[1] https://archive.ics.uci.edu/ml/datasets/auto+mpg

3. Korelasyon Analizi

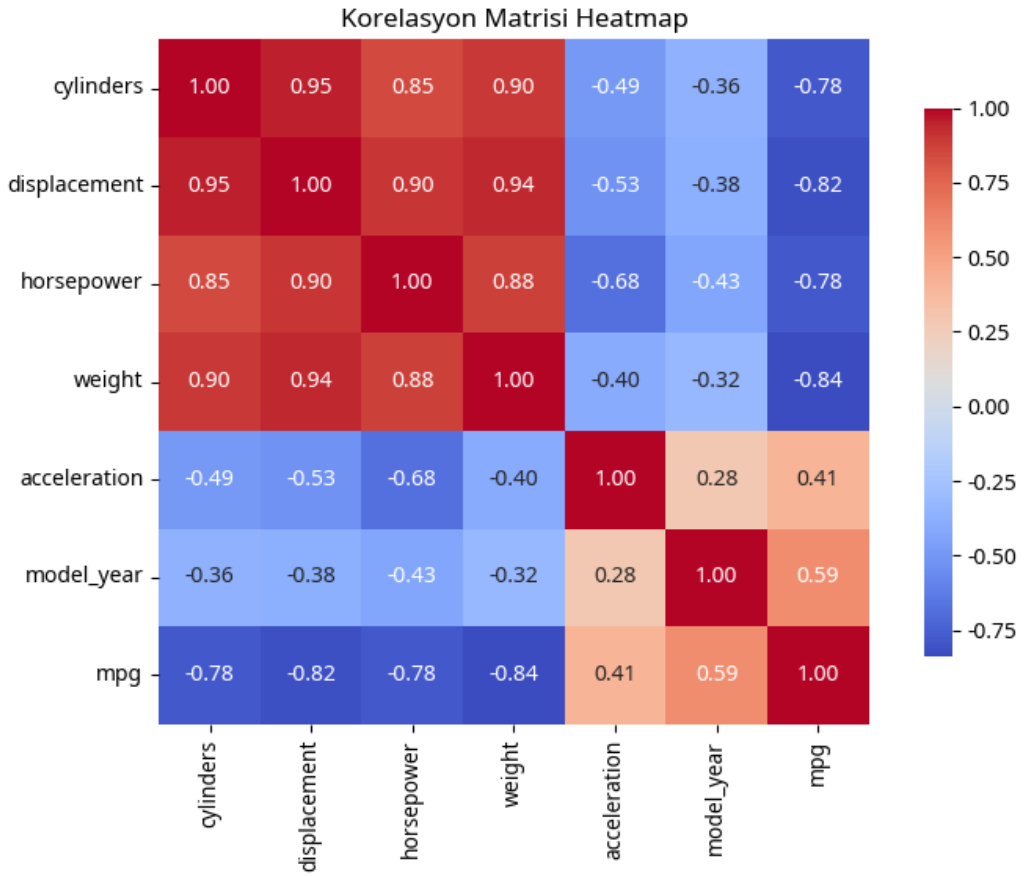
Basit doğrusal regresyon modelinde, mpg bağımlı değişkeni ile bağımsız değişkenler arasındaki ilişkinin gücünü ve yönünü belirlemek amacıyla Pearson Korelasyon Analizi yapılmıştır. Eğitim veri seti üzerinde yapılan analizler sonucunda, weight ile mpg arasında -0.84 gibi oldukça güçlü ve negatif yönlü bir korelasyon katsayısı bulunmuştur. Bu güçlü ilişki nedeniyle, modelde bağımsız değişken olarak ağırlık (weight) değişkeni özellikle tercih edilmiştir. Bu durum, aracın ağırlığı arttıkça yakıt tüketiminin arttığını, yani aracın daha fazla yakıt harcadığını göstermektedir. Korelasyon katsayısının p-değeri 0.0000 olarak hesaplanmış olup, bu değer istatistiksel olarak anlamlı bir ilişkinin varlığını göstermektedir ($p < 0.05$). Diğer bağımsız değişkenler ile mpg arasındaki Pearson korelasyon katsayıları ve p-değerleri aşağıda özetlenmiştir:

- cylinders: -0.78 (p-değeri: 0.0000) - İstatistiksel olarak anlamlı
- displacement: -0.82 (p-değeri: 0.0000) - İstatistiksel olarak anlamlı
- horsepower: -0.78 (p-değeri: 0.0000) - İstatistiksel olarak anlamlı
- weight: -0.84 (p-değeri: 0.0000) - İstatistiksel olarak anlamlı
- acceleration: 0.41 (p-değeri: 0.0000) - İstatistiksel olarak anlamlı
- model_year: 0.59 (p-değeri: 0.0000) - İstatistiksel olarak anlamlı

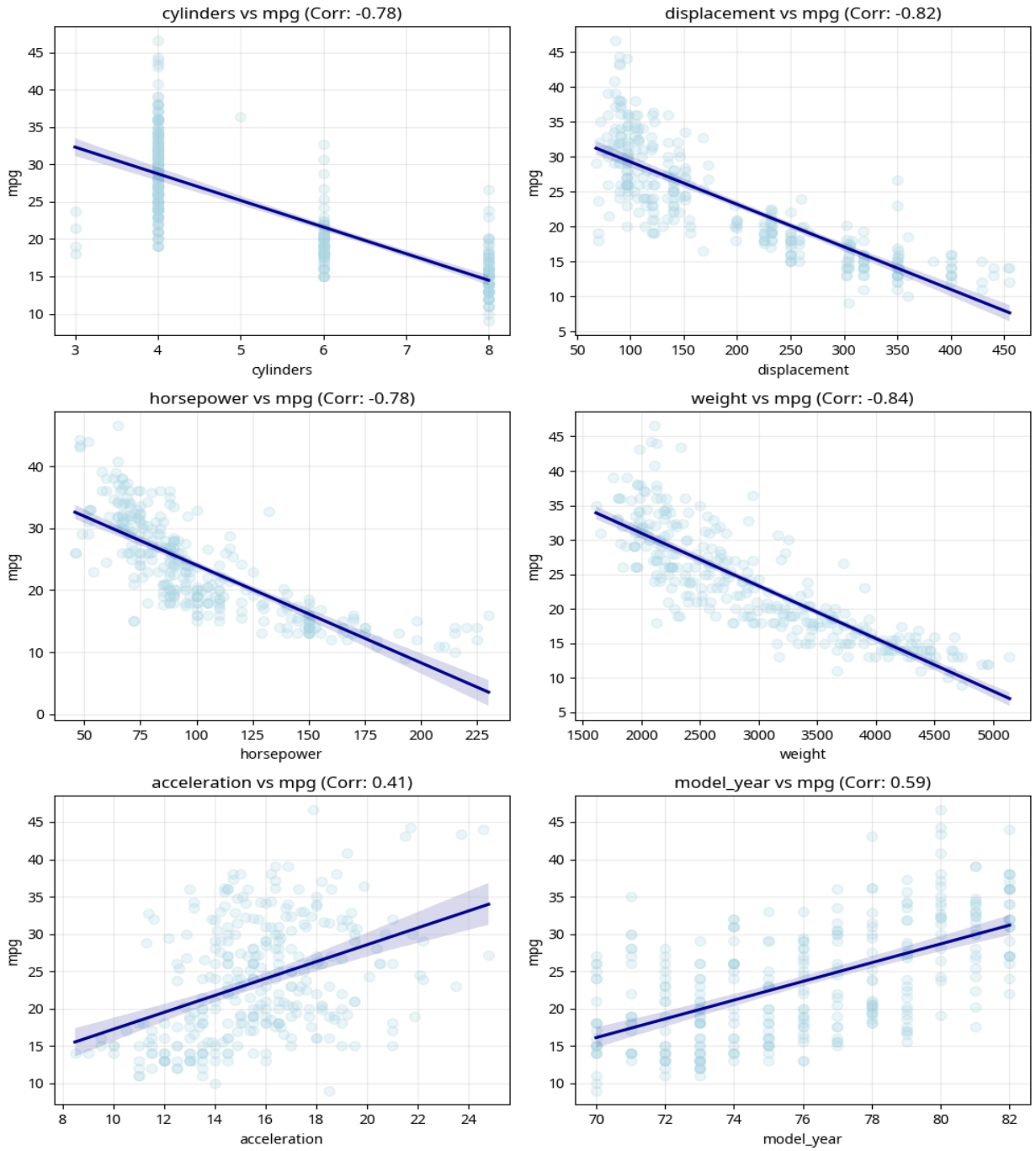
Bu sonuçlar, weight değişkeninin mpg üzerindeki etkisinin en güçlü negatif ilişkiye sahip olduğunu doğrulamaktadır. Tüm değişkenlerin mpg ile istatistiksel olarak anlamlı bir korelasyona sahip olduğu görülmüştür.

Görselleştirmeler:

Korelasyon matrisi heatmap ve değişkenler arası dağılım grafikleri (scatter plot matrisi) ile ilişkiler görsel olarak incelenmiştir. Korelasyon heatmap, değişkenler arasındaki ikili ilişkileri renk yoğunluğu ile gösterirken, scatter plot matrisi her bir bağımsız değişken ile mpg arasındaki ilişkinin dağılımını ve regresyon doğrusunu sunmaktadır.



Şekil 1, veri setindeki tüm sayısal değişkenler arasındaki korelasyon katsayılarını göstermektedir. mpg ile weight arasındaki güçlü negatif ilişki (-0.84) açıkça görülmektedir.



Şekil 2, her bir bağımsız değişkenin mpg ile olan ilişkisini gösteren dağılım grafiklerini sunmaktadır. Özellikle weight ile mpg arasındaki doğrusal negatif ilişki bu grafikte belirgin bir şekilde gözlemlenmektedir.

4. Temel Model (Orijinal Veri)

Bu bölümde, herhangi bir veri dönüşümü uygulanmadan, orijinal weight ve mpg değişkenleri kullanılarak oluşturulan basit doğrusal regresyon modeli incelenmiştir. Model, mpg'yi weight'in bir fonksiyonu olarak tahmin etmektedir. Eğitim veri seti üzerinde kurulan modelin istatistiksel özeti aşağıda sunulmuştur:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.699
Model:                            OLS   Adj. R-squared:                  0.698
Method:                 Least Squares   F-statistic:                     720.8
Date:                Mon, 21 Jul 2025   Prob (F-statistic):           5.54e-83
Time:                        17:07:09   Log-Likelihood:                -899.13
No. Observations:                 313   AIC:                             1802.
Df Residuals:                     311   BIC:                             1810.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         46.2212      0.876     52.763      0.000      44.498      47.945
weight        -0.0076      0.000    -26.847      0.000      -0.008      -0.007
==============================================================================
Omnibus:                       29.271   Durbin-Watson:                   2.040
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               41.469
Skew:                           0.639   Prob(JB):                     9.89e-10
Kurtosis:                       4.243   Cond. No.                     1.11e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 1.11e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model Denklemi:

Modelin denklemi şu şekildedir:

mpg = 46.2212 - 0.0076 * weight

Bu denklem, weight değişkenindeki her bir birimlik artışın mpg değerinde 0.0076 birimlik bir azalmaya neden olduğunu göstermektedir. const (sabit) terimi ise weight sıfır olduğunda mpg'nin tahmini değerini ifade eder.

Performans Metrikleri:

1. *R-kare (R^2):*
   Modelin R^2 değeri 0.699 olarak bulunmuştur. Bu, mpg değişkenindeki varyansın yaklaşık %69.9'unun weight değişkeni tarafından açıklandığı anlamına gelmektedir. Bu değer, modelin bağımlı değişkendeki değişimi açıklama gücünün orta düzeyde olduğunu göstermektedir.

2. *Ortalama Mutlak Hata (MAE):*

   Eğitim seti üzerinde hesaplanan MAE değeri 3.230'dur. Bu, modelin tahminlerinin gerçek mpg değerlerinden ortalama olarak 3.230 birim saptığını ifade eder.

Varsayım Testleri:

Doğrusal regresyon modelinin geçerliliği için artıkların (residuals) belirli varsayımları karşılaması gerekmektedir: normallik ve homoskedastisite (sabit varyanslılık).

1. *Shapiro-Wilk Normallik Testi:*
   Artıkların normal dağılıp dağılmadığını test etmek için Shapiro-Wilk testi uygulanmıştır. Test sonucu (W İstatistiği = 0.97, p değeri = 0.00002) p-değerinin (0.00002) anlamlılık düzeyi olan 0.05'ten küçük olduğunu göstermektedir. Bu durum, istatistiksel olarak artıkların normal dağıldığı anlamlı bir şekilde söylenemez sonucunu ortaya koymaktadır. Yani, artıklar normal dağılım varsayımını ihlal etmektedir.

2. *Breusch-Pagan Homoskedastisite Testi:*
   Artıkların sabit varyansa sahip olup olmadığını test etmek için Breusch-Pagan testi uygulanmıştır. Test sonucu (p değeri = 0.00003) p-değerinin (0.00003) anlamlılık düzeyi olan 0.05'ten küçük olduğunu göstermektedir. Bu durum, istatistiksel olarak artıkların sabit varyansa sahip olduğu anlamlı bir şekilde söylenemez sonucunu ortaya koymaktadır. Yani, artıklar homoskedastisite varsayımını ihlal etmektedir.

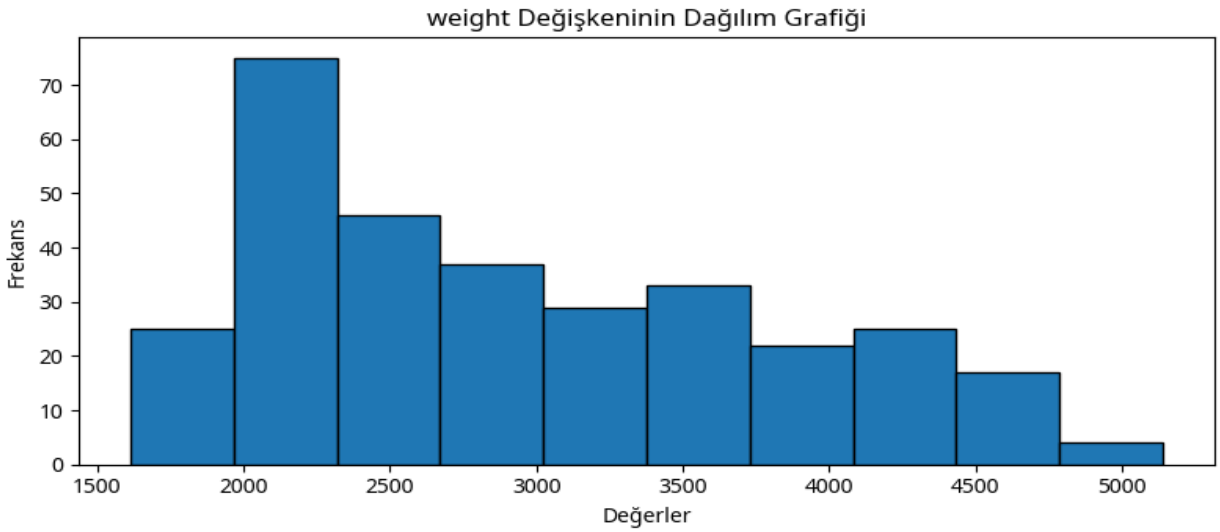
3. *Durbin-Watson Otokorelasyon Testi:*
   Artıklar arasında otokorelasyon (bağımlılık) olup olmadığını test etmek amacıyla Durbin-Watson testi uygulanmıştır. Test sonucu 2.040 olarak bulunmuştur. Durbin-Watson değeri 2'ye oldukça yakın olduğundan, artıklar arasında anlamlı bir otokorelasyon bulunmadığı, yani artıkların birbirinden bağımsız olduğu söylenebilir. Bu sonuç, modelin bağımsızlık varsayımını sağladığını göstermektedir.

Her üç varsayım testinin sonuçları birlikte değerlendirildiğinde, Shapiro-Wilk ve Breusch-Pagan testlerinde elde edilen p-değerlerinin 0.05'ten küçük olması, modelin artıklarının normal dağılmadığını ve sabit varyansa sahip olmadığını göstermektedir. Bu durum, modelin tahminlerinin güvenilirliğini ve geçerliliğini olumsuz etkileyebilir. Buna karşın, Durbin-Watson testinde elde edilen 2.040 değeri, artıklar arasında anlamlı bir otokorelasyon olmadığını, yani bağımsızlık varsayımının sağlandığını göstermektedir.

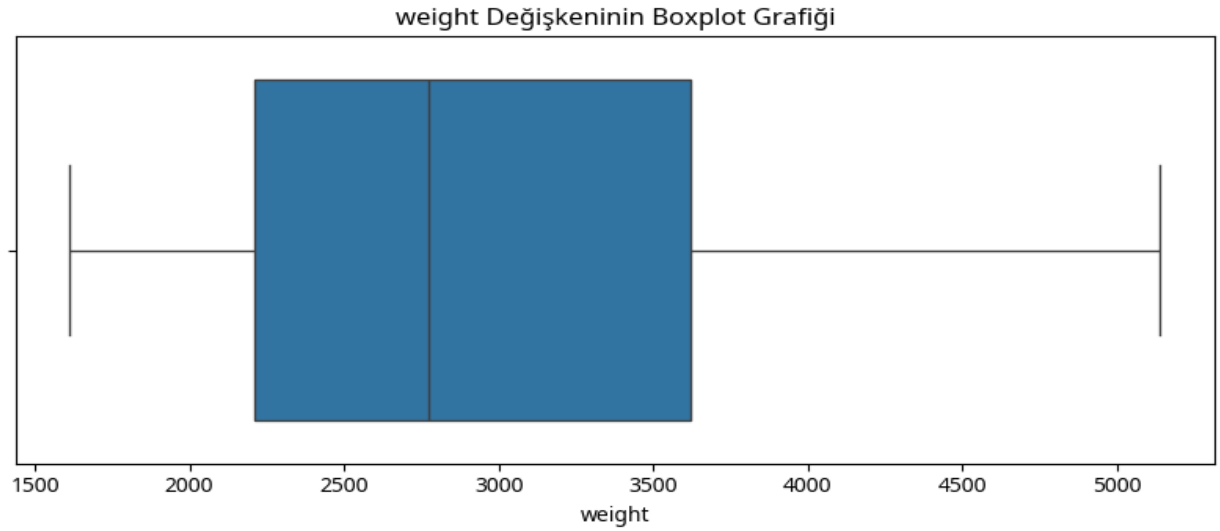
Görselleştirmeler:

Değişken Dağılım Grafikleri:

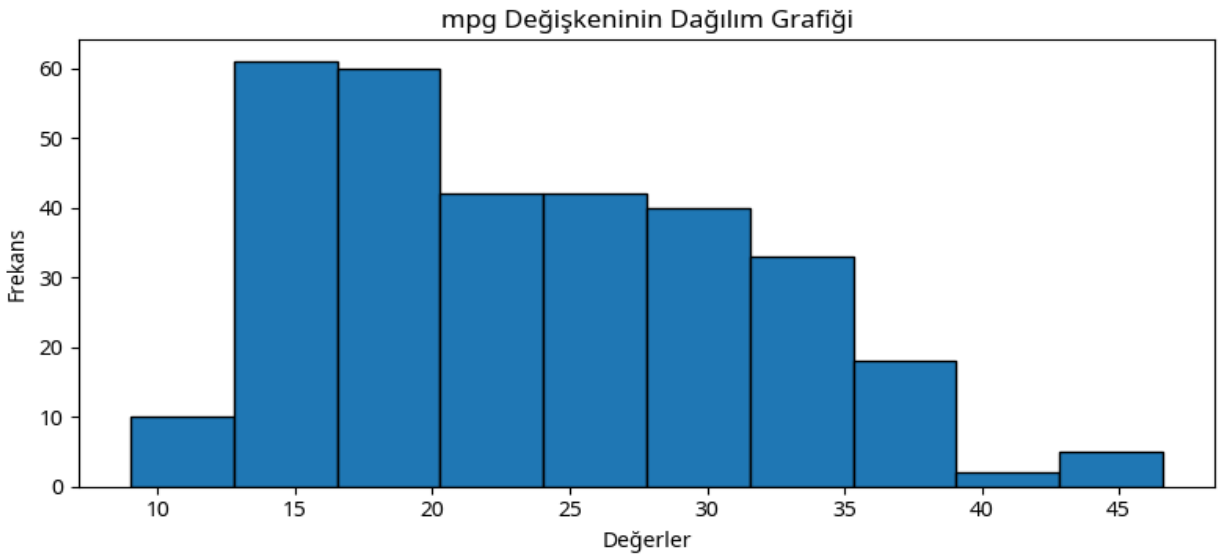
weight ve mpg değişkenlerinin orijinal dağılımları histogram ve kutu grafikleri ile incelenmiştir. Bu grafikler, değişkenlerin çarpıklıklarını ve aykırı değerlerini görselleştirmeye yardımcı olur.



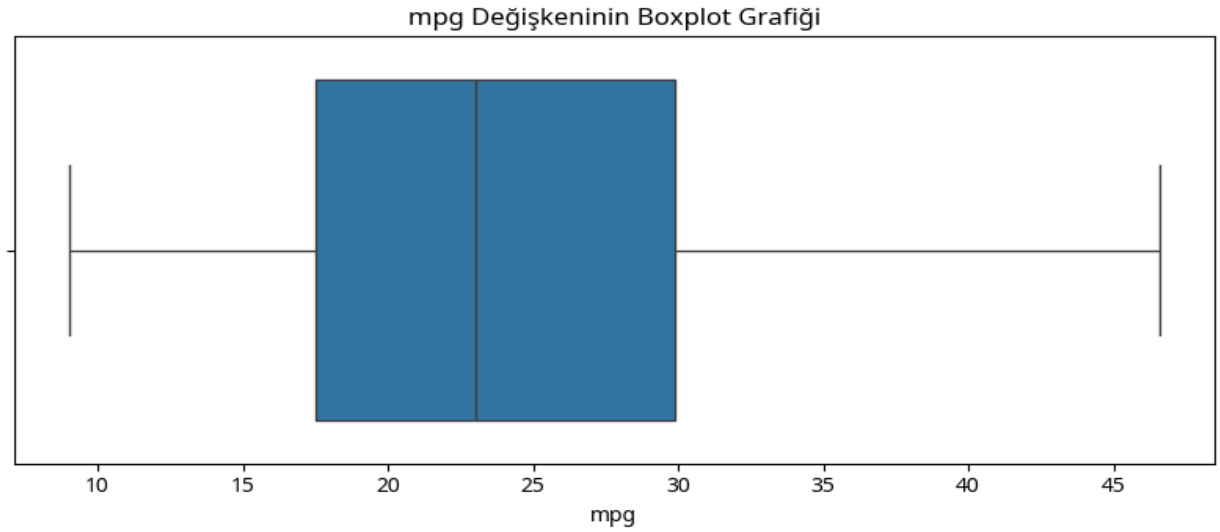
Şekil 3, weight değişkeninin dağılımını göstermektedir. Dağılımın sağa çarpık olduğu gözlemlenmektedir.



Şekil 4, weight değişkeninin dağılımını göstermektedir. Dağılımın sağa çarpık olduğu ve aykırı değer içermediği gözlemlenmektedir.



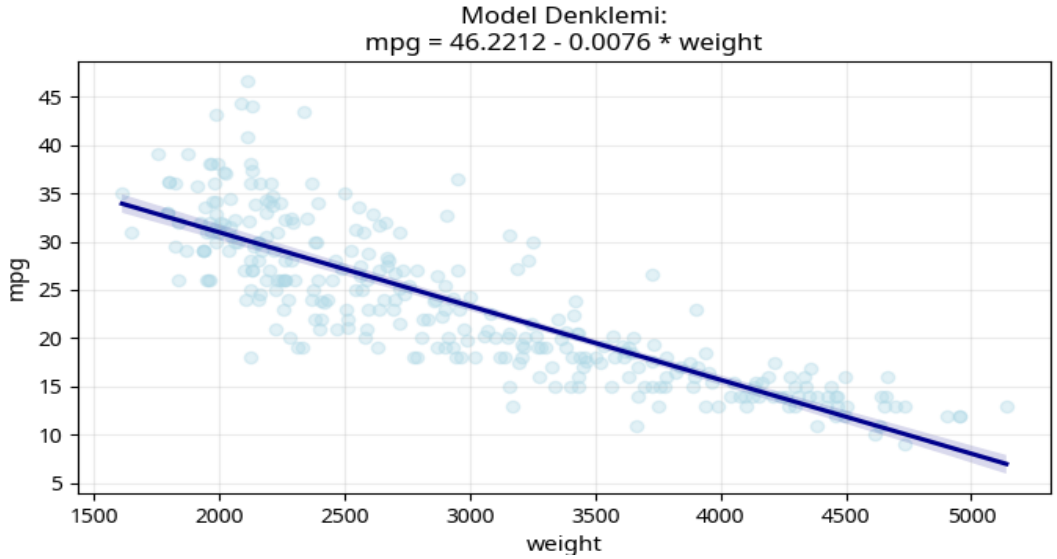
Şekil 5, mpg değişkeninin dağılımını göstermektedir. Dağılımın sağa çarpık olduğu gözlemlenmektedir.



Şekil 6, mpg değişkeninin dağılımını göstermektedir. Dağılımın sağa çarpık olduğu ve aykırı değer içermediği gözlemlenmektedir.

Regresyon Doğrusu:

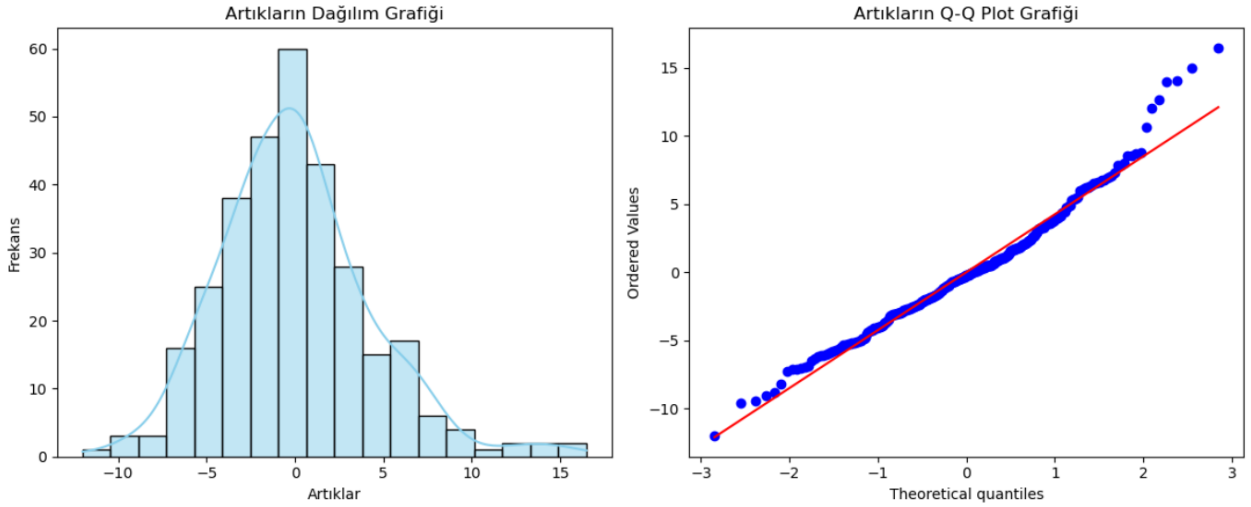
Orijinal veriler üzerinde oluşturulan regresyon doğrusu, weight ile mpg arasındaki ilişkiyi görsel olarak sunmaktadır.



Şekil 7, weight ve mpg arasındaki ilişkiyi ve tahmin edilen regresyon doğrusunu göstermektedir. Negatif eğim, ağırlık arttıkça yakıt tüketiminin azaldığını doğrulamaktadır.

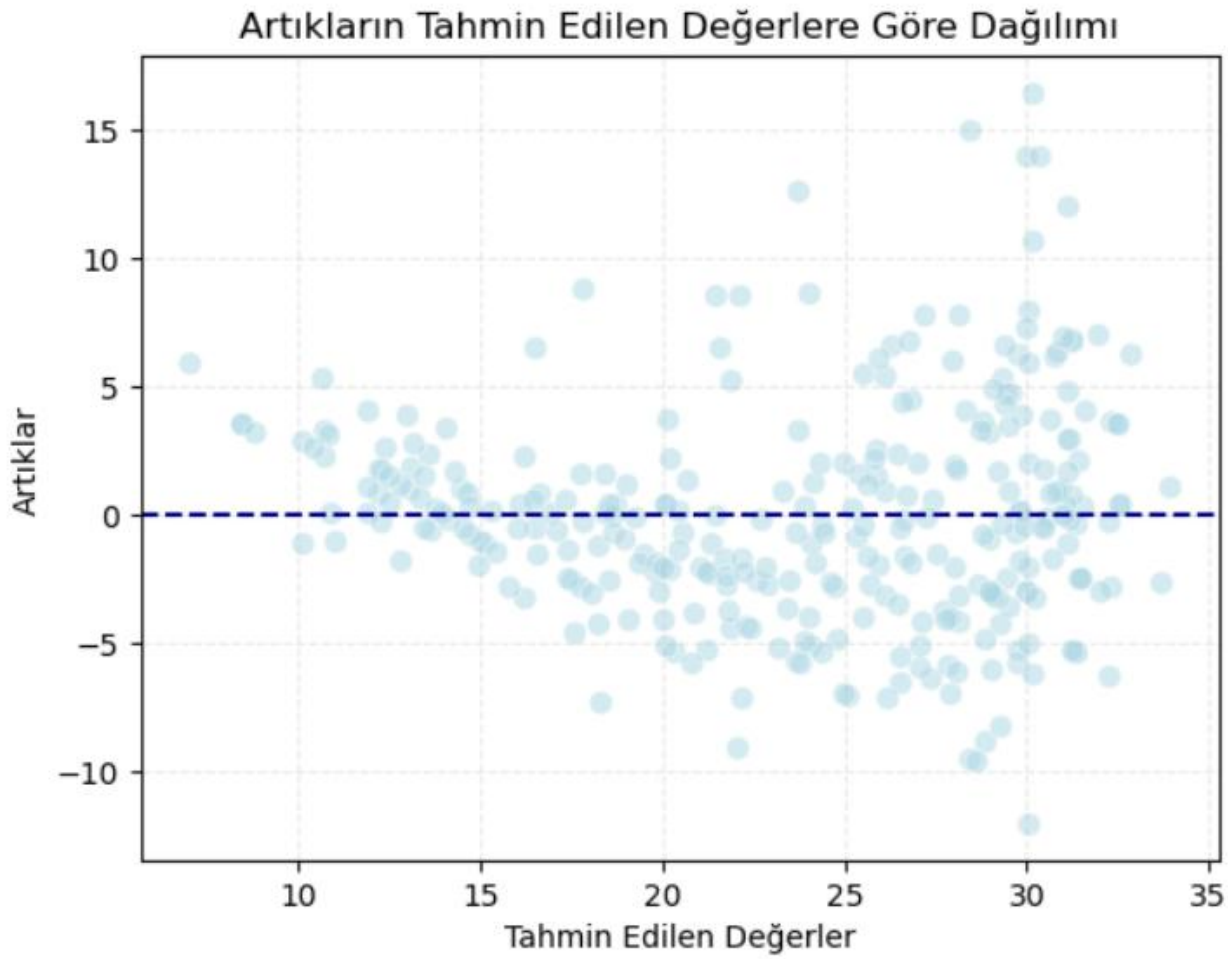
Artık Analizi Grafikleri:

Artıkların dağılımı ve tahmin edilen değerlere göre dağılımı, varsayım ihlallerini görsel olarak doğrulamak için kullanılmıştır.



Artıkların Dağılım Grafiği (Orijinal Model) Şekil 8, orijinal modelin artıklarının dağılımını göstermektedir.

Artıkların Q-Q Plot Grafiği (Orijinal Model) Şekil 9, orijinal modelin artıklarının Q-Q plotunu göstermektedir. Q-Q plot üzerindeki noktaların normal dağılım çizgisinden sapması, normallik varsayımının ihlal edildiğini desteklemektedir.



Şekil 10, artıkların tahmin edilen değerlere göre dağılımını göstermektedir. Artıkların rastgele dağılmayıp belirli bir patern (huni şekli) sergilemesi, homoskedastisite varsayımının ihlal edildiğini (heteroskedastisite) açıkça ortaya koymaktadır.

5. Veri Dönüşümü Analizi

Önceki bölümde görüldüğü üzere, orijinal verilerle kurulan basit doğrusal regresyon modeli, normallik ve homoskedastisite varsayımlarını ihlal etmektedir. Bu varsayım ihlalleri, modelin tahminlerinin güvenilirliğini ve istatistiksel çıkarımların geçerliliğini olumsuz etkileyebilir. Bu nedenle, model varsayımlarını sağlamak ve model performansını iyileştirmek amacıyla weight (bağımsız değişken) ve mpg (bağımlı değişken) üzerinde çeşitli veri dönüşümleri denenmiştir.

Her bir değişken için altı farklı dönüşüm türü (Original, Log, Sqrt, Reciprocal, Box-Cox, Square) değerlendirilmiştir. Bu, toplamda 36 farklı dönüşüm kombinasyonunun test edildiği anlamına gelmektedir. Her bir kombinasyon için yeni bir regresyon modeli kurulmuş ve artıkların normallik (Shapiro-Wilk testi) ve homoskedastisite (Breusch-Pagan testi) varsayımları kontrol edilmiştir. Ayrıca, her modelin açıklayıcılık gücü R-kare (R^2) değeri ile ölçülmüştür.

Dönüşüm Kombinasyonlarının Değerlendirilmesi:

Değerlendirme sürecinde, öncelikli olarak hem normallik hem de homoskedastisite varsayımlarını (p-değeri > 0.05) sağlayan kombinasyonlar aranmıştır. Bu koşulu sağlayan kombinasyonlar arasında en yüksek R^2 değerine sahip olan tercih edilmiştir. Eğer hiçbir kombinasyon her iki varsayımı da sağlamazsa, varsayımlardan en az birini sağlayan ve R^2 değeri yüksek olan kombinasyonlar değerlendirilmiştir.

Analiz sonuçlarına göre, Sqrt(weight) ve Log(mpg) dönüşüm kombinasyonu, her iki varsayımı da istatistiksel olarak anlamlı bir şekilde sağlayan ve aynı zamanda yüksek bir R^2 değerine sahip olan en uygun seçim olarak belirlenmiştir. Bu kombinasyon için elde edilen metrikler aşağıdaki gibidir:

X Transformation	Y Transformation	R^2	Shapiro Wilk p	Breusch-Pagan p
Sqrt	Log	0.7720	0.0593	0.1159

Bu tablo, Sqrt(weight) ve Log(mpg) dönüşümlerinin, artıkların normal dağılım (Shapiro-Wilk p-değeri = 0.0593 > 0.05) ve sabit varyans (Breusch-Pagan p-değeri = 0.1159 > 0.05) varsayımlarını sağladığını göstermektedir. Ayrıca, bu dönüşümlerle modelin R^2 değeri 0.699'dan 0.772'ye yükselmiştir, bu da modelin açıklayıcılık gücünde önemli bir iyileşme olduğunu göstermektedir.

Diğer dikkat çekici kombinasyonlar arasında Original(weight) ve Log(mpg) (R^2=0.7703) ile Square(weight) ve Log(mpg) (R^2=0.7482) yer almaktadır. Bu kombinasyonlar da varsayımları sağlamış ve yüksek R^2 değerleri sunmuştur. Ancak, en yüksek R^2 değeri ve varsayım sağlama kriterleri göz önüne alındığında Sqrt(weight) ve Log(mpg) kombinasyonu optimum seçim olarak öne çıkmıştır.

Varsayımlardan bağımsız olarak en yüksek R^2 değerine sahip ilk 3 kombinasyon ise şunlardır:

X Transformation	Y Transformation	R^2	Shapiro Wilk p	Breusch-Pagan p
Square	Reciprocal	0.7875	0.0000	0.0006
Original	Reciprocal	0.7863	0.0000	0.0014
Sqrt	Reciprocal	0.7760	0.0000	0.0017

Bu kombinasyonlar yüksek R^2 değerleri sunsa da, artıkların normallik ve homoskedastisite varsayımlarını sağlamadığı (p-değerleri < 0.05) için tercih edilmemiştir. Regresyon analizinde varsayımların sağlanması, modelin istatistiksel geçerliliği açısından kritik öneme sahiptir.

Metodolojik Seçimlerin Gerekçelendirilmesi:

Veri dönüşümü seçimi, sadece R^2 değerini maksimize etmekle kalmayıp, aynı zamanda regresyon modelinin temel varsayımlarını karşılamasını sağlamayı amaçlamıştır. Varsayımların ihlali, katsayı tahminlerinin yanlı olmasına, standart hataların yanlış hesaplanmasına ve dolayısıyla p-değerlerinin ve güven aralıklarının güvenilmez olmasına yol açabilir. Bu nedenle, Sqrt(weight) ve Log(mpg) dönüşümlerinin seçimi, hem modelin açıklayıcılık gücünü artırması hem de istatistiksel geçerliliğini sağlaması açısından uygun bulunmuştur.

6. Dönüştürülmüş Model

Önceki bölümde yapılan dönüşüm analizi sonucunda, weight değişkeni için karekök (Sqrt) dönüşümü ve mpg değişkeni için logaritmik (Log) dönüşümünün en uygun kombinasyon olduğu belirlenmiştir. Bu dönüşümler uygulandıktan sonra yeni bir basit doğrusal regresyon modeli oluşturulmuştur. Modelin istatistiksel özeti aşağıda sunulmuştur:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.772
Model:                            OLS   Adj. R-squared:                  0.771
Method:                 Least Squares   F-statistic:                     1053.
Date:                Mon, 21 Jul 2025   Prob (F-statistic):          7.24e-102
Time:                        17:07:11   Log-Likelihood:                 125.49
No. Observations:                 313   AIC:                            -247.0
Df Residuals:                     311   BIC:                            -239.5
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          5.1896      0.065     80.021      0.000       5.062       5.317
x1            -0.0387      0.001    -32.454      0.000      -0.041      -0.036
==============================================================================
Omnibus:                        5.372   Durbin-Watson:                   2.027
Prob(Omnibus):                  0.068   Jarque-Bera (JB):                6.308
Skew:                           0.160   Prob(JB):                       0.0427
Kurtosis:                       3.617   Cond. No.                         384.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
```

Dönüştürülmüş Model Denklemi:

Modelin denklemi şu şekildedir:

log(mpg) = 5.1896 - 0.0387 * sqrt(weight)

Bu denklem, sqrt(weight) değişkenindeki her bir birimlik artışın log(mpg) değerinde 0.0387 birimlik bir azalmaya neden olduğunu göstermektedir.

İyileşmiş Performans Metrikleri:

1. *R-kare (R^2):*

   Dönüştürülmüş modelin R^2 değeri 0.772 olarak bulunmuştur. Orijinal modelin R^2 değeri 0.699 iken, bu önemli artış (%7.3), dönüşümlerin modelin açıklayıcılık gücünü kayda değer ölçüde iyileştirdiğini göstermektedir. Bu, log(mpg) değişkenindeki varyansın yaklaşık %77.2%sinin sqrt(weight) değişkeni tarafından açıklandığı anlamına gelmektedir.

2. *Ortalama Mutlak Hata (MAE):*
   Eğitim seti üzerinde hesaplanan MAE değeri, orijinal ölçeğe geri dönüştürüldüğünde 3.016 olarak bulunmuştur. Orijinal modelin MAE değeri 3.230 iken, bu düşüş, dönüştürülmüş modelin tahminlerinin gerçek mpg değerlerine daha yakın olduğunu ve tahmin hatasının azaldığını göstermektedir.

Varsayım Testleri (Dönüştürülmüş Model):

Dönüştürülmüş modelin artıklarının varsayım testleri aşağıdaki gibidir:

1. *Shapiro-Wilk Normallik Testi:*
   Test sonucu (W İstatistiği = 0.99, p değeri = 0.05932) p-değerinin (0.05932) anlamlılık düzeyi olan 0.05%ten büyük olduğunu göstermektedir. Bu durum, istatistiksel olarak artıkların normal dağıldığı anlamlı bir şekilde söylenebilir sonucunu ortaya koymaktadır. Böylece normallik varsayımı sağlanmıştır.

2. *Breusch-Pagan Homoskedastisite Testi:*

   Test sonucu (p değeri = 0.11591) p-değerinin (0.11591) anlamlılık düzeyi olan 0.05%ten büyük olduğunu göstermektedir. Bu durum, istatistiksel olarak artıkların sabit varyansa sahip olduğu anlamlı bir şekilde söylenebilir sonucunu ortaya koymaktadır. Böylece homoskedastisite varsayımı da sağlanmıştır.

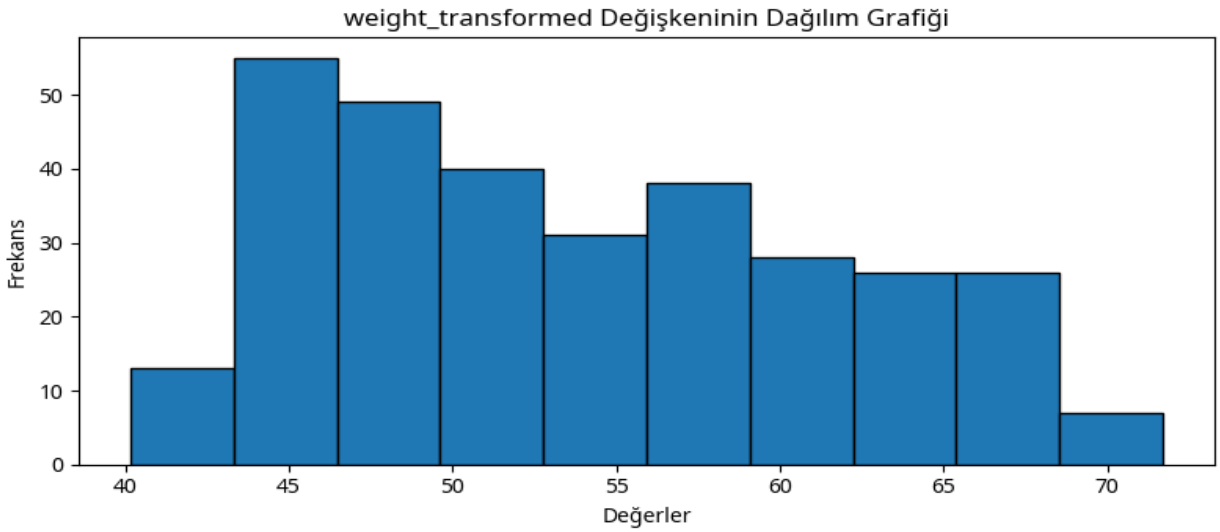
3. *Durbin-Watson Otokorelasyon Testi:*

   Artıklar arasında otokorelasyon (bağımlılık) olup olmadığını test etmek amacıyla Durbin-Watson testi uygulanmıştır. Test sonucu 2.027 olarak bulunmuştur. Bu değer 2'ye oldukça yakın olduğundan, artıklar arasında anlamlı bir otokorelasyon bulunmadığı, yani artıkların birbirinden bağımsız olduğu söylenebilir. Böylece bağımsızlık varsayımı da sağlanmıştır.

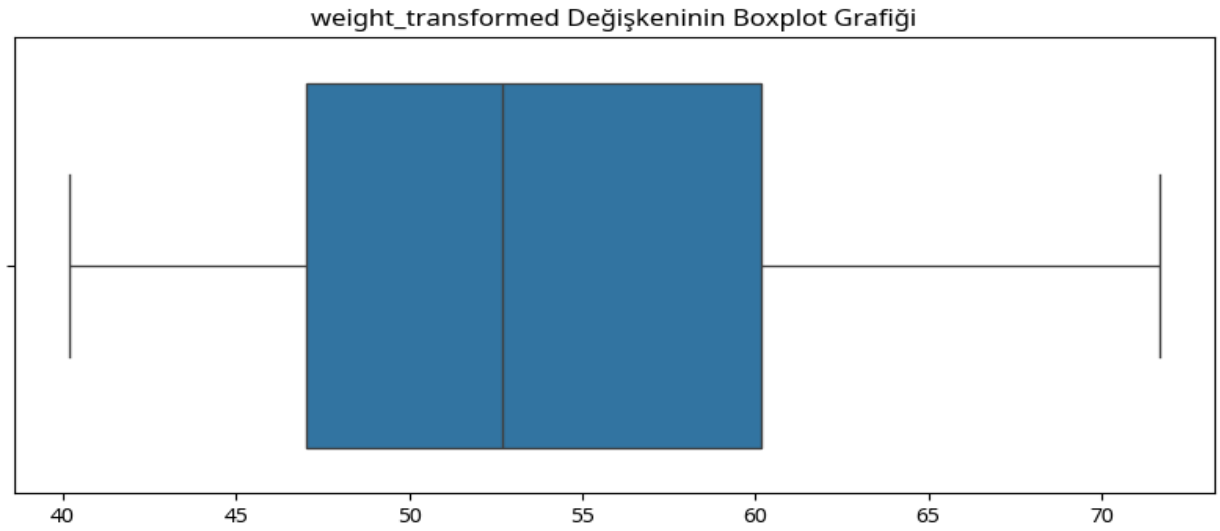
Bu sonuçlar, uygulanan veri dönüşümlerinin modelin varsayımlarını başarılı bir şekilde karşılamasını sağladığını göstermektedir. Shapiro-Wilk testi ile normallik varsayımı, Breusch-Pagan testi ile homoskedastisite (sabit varyans) varsayımı ve Durbin-Watson testi ile artıkların bağımsızlığı varsayımı sağlanmıştır.

Görselleştirmeler:

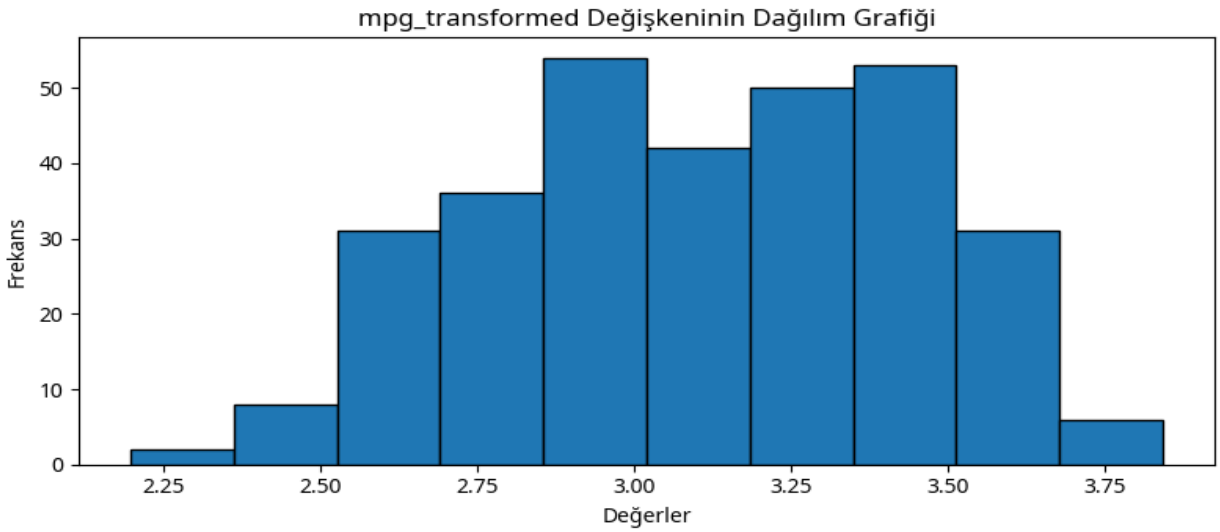
Dönüştürülmüş Değişken Dağılım Grafikleri:
sqrt(weight) ve log(mpg) değişkenlerinin dağılımları histogram ve kutu grafikleri ile incelenmiştir. Bu grafikler, dönüşümlerin değişkenlerin dağılımlarını nasıl normalleştirdiğini görselleştirmeye yardımcı olur.



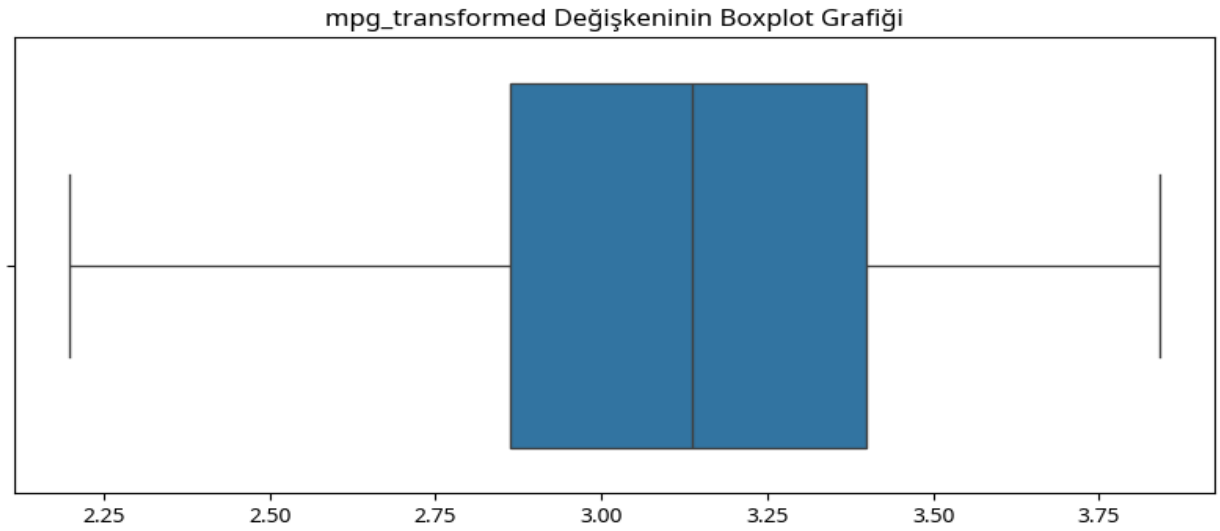
Şekil 11, sqrt(weight) değişkeninin dağılımını göstermektedir. Orijinal weight dağılımına göre daha simetrik bir yapıya sahip olduğu gözlemlenmektedir.



Şekil 12, sqrt(weight) değişkeninin kutu grafiğini sunmaktadır. Orijinal weight dağılımına göre daha simetrik bir yapıya sahip olduğu gözlemlenmektedir.



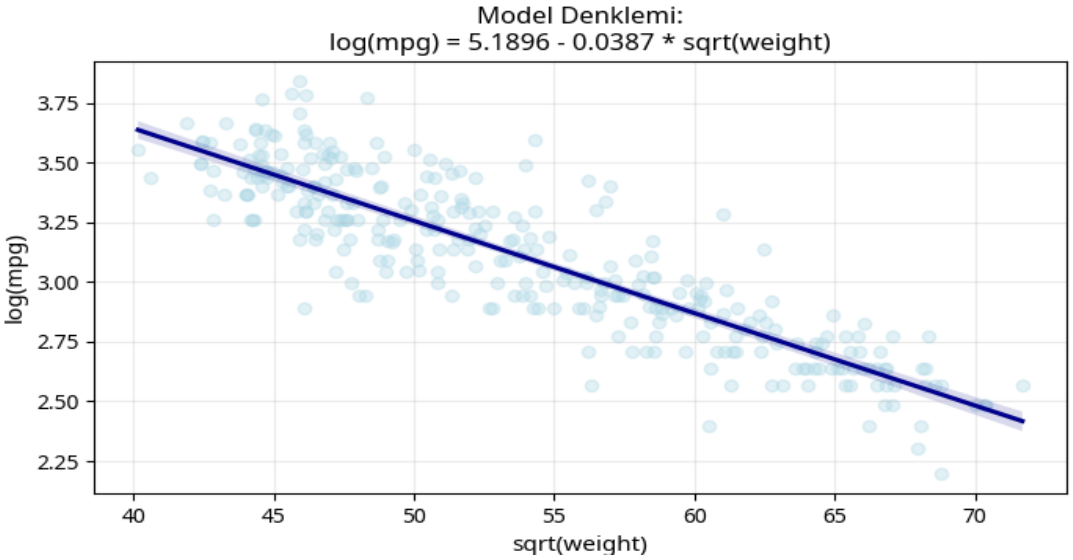
Şekil 13, log(mpg) değişkeninin dağılımını göstermektedir. Orijinal mpg dağılımına göre daha normal bir dağılıma yaklaştığı gözlemlenmektedir.



Şekil 14, log(mpg) değişkeninin kutu grafiğini sunmaktadır. Orijinal mpg dağılımına göre daha normal bir dağılıma yaklaştığı gözlemlenmektedir.

Dönüştürülmüş Regresyon Doğrusu:

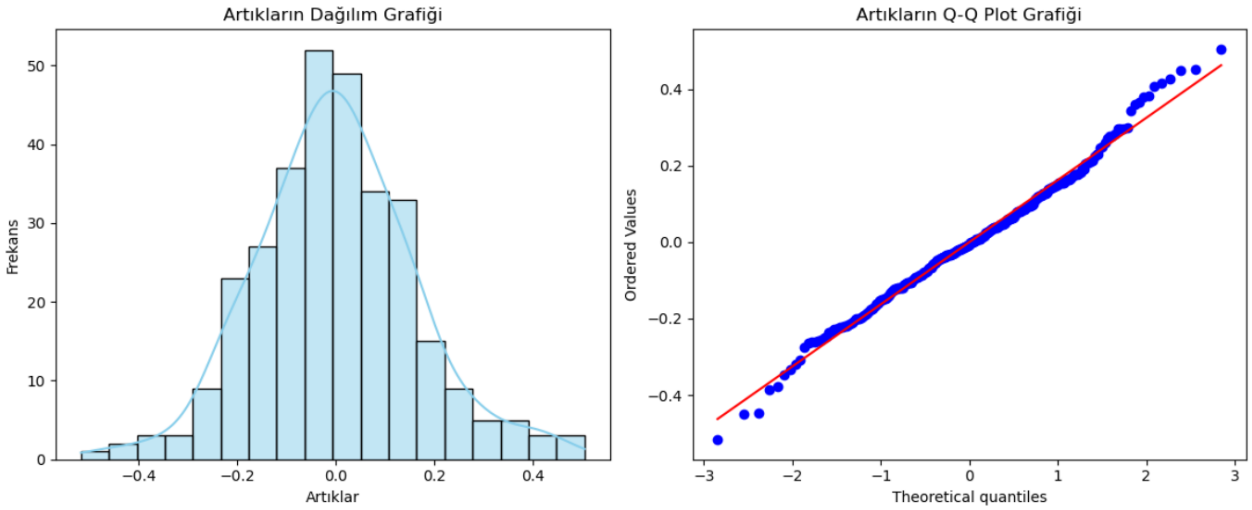
Dönüştürülmüş veriler üzerinde oluşturulan regresyon doğrusu, sqrt(weight) ile log(mpg) arasındaki ilişkiyi görsel olarak sunmaktadır.



Şekil 15, sqrt(weight) ve log(mpg) arasındaki ilişkiyi ve tahmin edilen regresyon doğrusunu göstermektedir. Dönüşümler sayesinde daha doğrusal bir ilişki gözlemlenmektedir.

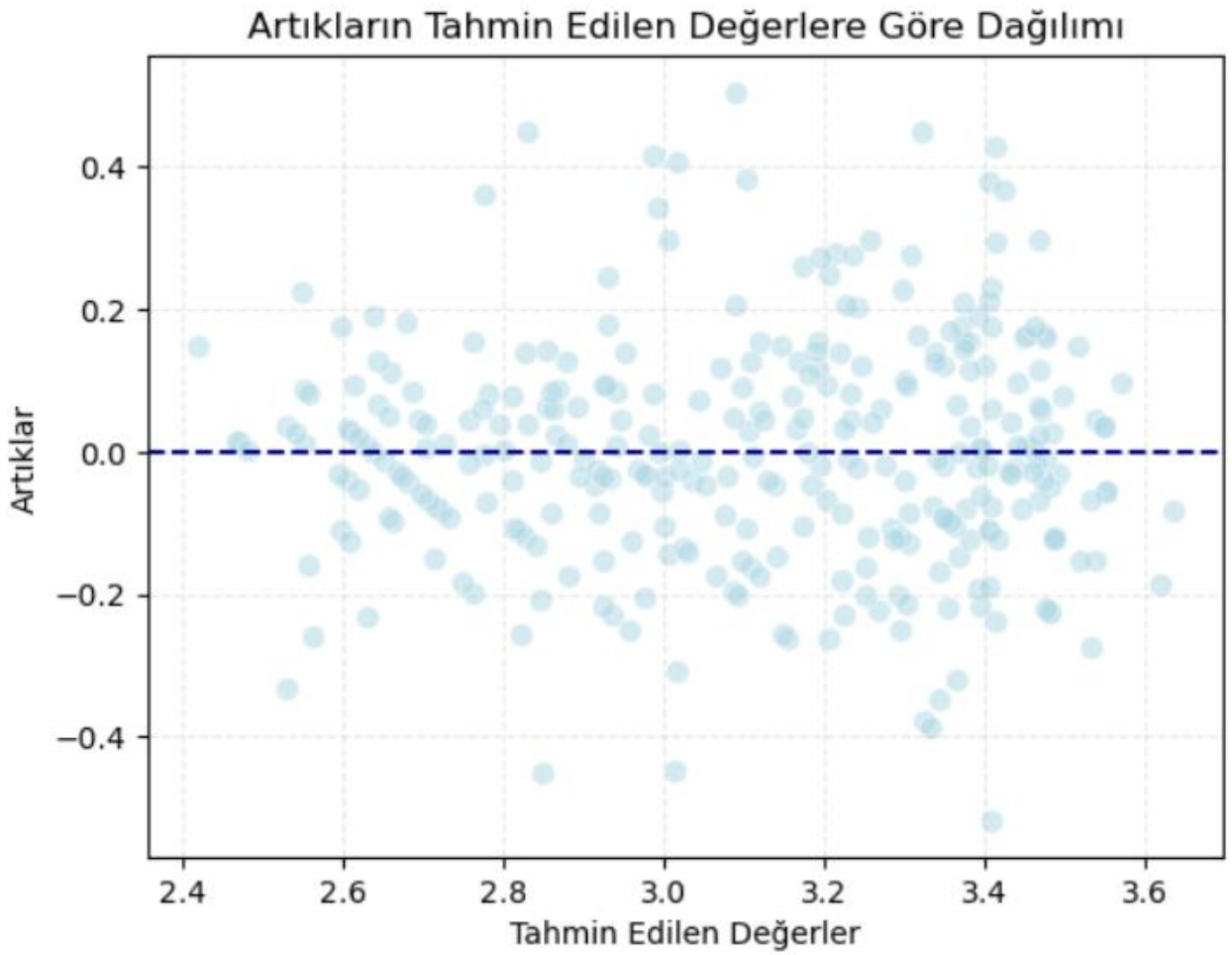
Artık Analizi Grafikleri (Dönüştürülmüş Model):

Dönüştürülmüş modelin artıklarının dağılımı ve tahmin edilen değerlere göre dağılımı, varsayımların sağlanıp sağlanmadığını görsel olarak doğrulamak için kullanılmıştır.



Artıkların Dağılım Grafiği (Dönüştürülmüş Model) Şekil 16, dönüştürülmüş modelin artıklarının dağılımını göstermektedir.

Artıkların Q-Q Plot Grafiği (Dönüştürülmüş Model) Şekil 17, dönüştürülmüş modelin artıklarının Q-Q plotunu göstermektedir. Q-Q plot üzerindeki noktaların normal dağılım çizgisine daha yakın olması, normallik varsayımının sağlandığını desteklemektedir.



Şekil 18, artıkların tahmin edilen değerlere göre dağılımını göstermektedir. Artıkların sıfır etrafında rastgele dağılması ve belirgin bir patern sergilememesi, homoskedastisite varsayımının sağlandığını açıkça ortaya koymaktadır.

7. Model Validasyonu

Dönüştürülmüş modelin genellenebilirliğini ve yeni, görünmeyen veriler üzerindeki performansını değerlendirmek amacıyla test veri seti kullanılarak model validasyonu yapılmıştır. Eğitim setinde belirlenen Sqrt(weight) ve Log(mpg) dönüşümleri, test setindeki weight ve mpg değişkenlerine de uygulanmıştır. Ardından, eğitim setinde eğitilen regresyon modeli, dönüştürülmüş test verileri üzerinde tahminler yapmak için kullanılmıştır.

Test Verisi Performansı:

1. *R-kare (R^2):*

   Test seti üzerinde hesaplanan R^2 değeri 0.698 olarak bulunmuştur. Eğitim setindeki R^2 değeri 0.772 iken, test setindeki bu değer, modelin genellenebilirliğinin eğitim setindeki kadar yüksek olmadığını ancak yine de kabul edilebilir bir açıklayıcılık gücüne sahip olduğunu göstermektedir. Eğitim ve test R^2 değerleri arasındaki fark, modelin eğitim verisine bir miktar aşırı uyum sağladığını (overfitting) düşündürebilir, ancak bu fark kritik düzeyde değildir.

2. *Ortalama Mutlak Hata (MAE):*

   Test seti üzerinde hesaplanan MAE değeri, orijinal ölçeğe geri dönüştürüldüğünde 3.088 olarak bulunmuştur. Eğitim setindeki MAE değeri 3.016 iken, test setindeki bu değer, modelin yeni veriler üzerinde de benzer bir tahmin hatası sergilediğini göstermektedir. Bu, modelin gerçek dünya senaryolarında da makul tahminler yapabileceğini düşündürmektedir.

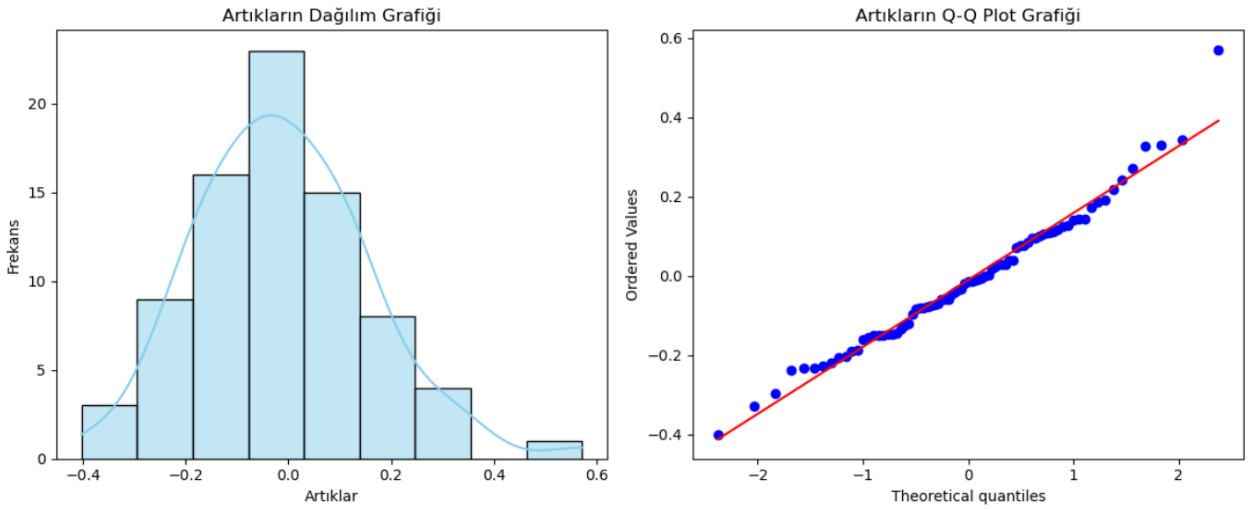
Ters Dönüşüm Sonuçları:

Modelin tahminleri log(mpg) ölçeğinde yapıldığı için, bu tahminlerin orijinal mpg ölçeğine geri dönüştürülmesi gerekmektedir. Bu işlem için np.exp() fonksiyonu kullanılarak ters logaritmik dönüşüm uygulanmıştır. Bu sayede, modelin gerçek mpg değerleri ile karşılaştırılabilecek tahminler elde edilmiştir.

Görselleştirmeler:

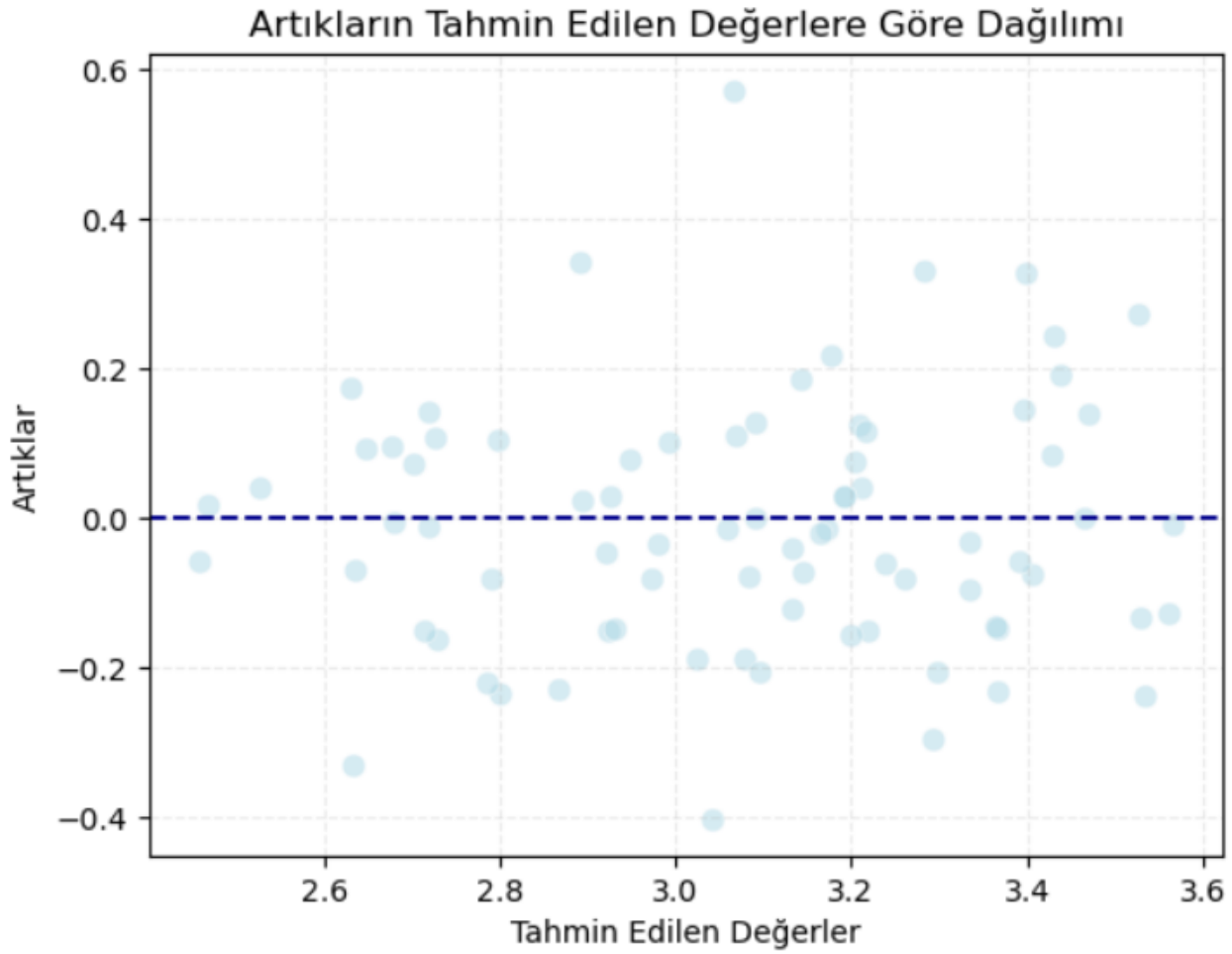
Test Artık Analizi Grafikleri:

Test setindeki artıkların dağılımı ve tahmin edilen değerlere göre dağılımı, modelin test verileri üzerindeki varsayım performansını değerlendirmek için kullanılmıştır.



Test Artıklarının Dağılım Grafiği Şekil 19, test setindeki artıkların dağılımını göstermektedir. Artıkların normal dağılıma yakın olduğu gözlemlenmektedir.

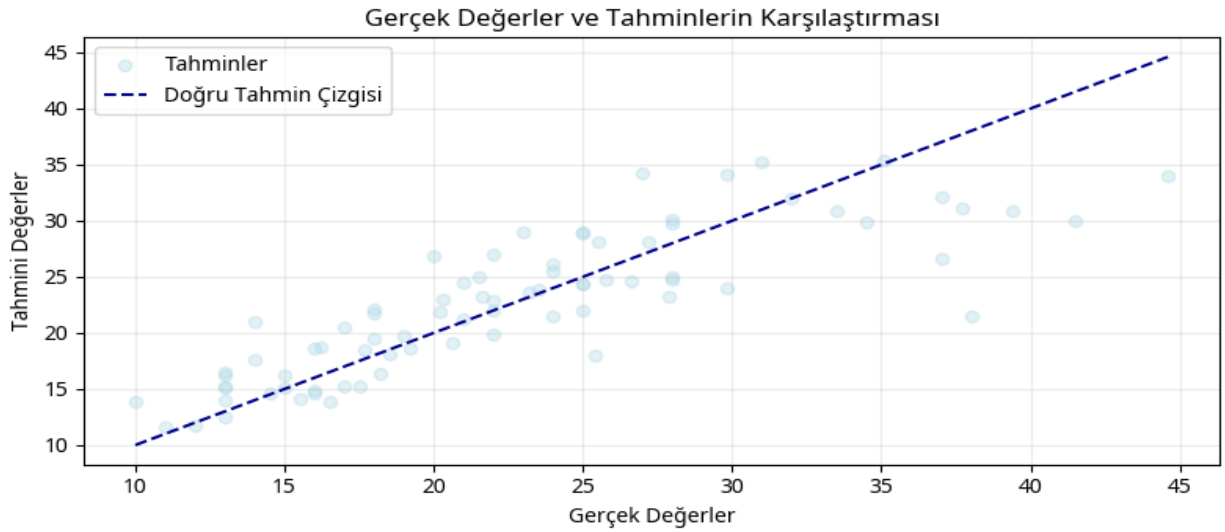
Test Artıklarının Q-Q Plot Grafiği Şekil 20, test setindeki artıkların Q-Q plotunu göstermektedir. Artıkların normal dağılıma yakın olduğu gözlemlenmektedir.



Şekil 21, test setindeki artıkların tahmin edilen değerlere göre dağılımını göstermektedir. Artıkların sıfır etrafında rastgele dağıldığı ve belirgin bir patern sergilemediği görülmektedir, bu da homoskedastisite varsayımının test setinde de korunduğunu düşündürmektedir.

Gerçek Değerler ve Tahminlerin Karşılaştırması:

Test setindeki gerçek mpg değerleri ile modelin tahmin ettiği mpg değerleri arasındaki ilişki bir dağılım grafiği ile görselleştirilmiştir.



Şekil 22, test setindeki gerçek mpg değerleri ile modelin tahmin ettiği mpg değerlerini karşılaştırmaktadır. Noktaların 45 derecelik doğru tahmin çizgisine yakın olması, modelin başarılı tahminler yaptığını göstermektedir. Özellikle düşük ve orta mpg değerlerinde tahminlerin daha isabetli olduğu, yüksek mpg değerlerinde ise bazı sapmaların olduğu gözlemlenmektedir.

8. Sonuçlar ve Değerlendirme

Bu proje, Auto MPG veri seti üzerinde weight değişkeni kullanılarak araç yakıt tüketimini (mpg) tahmin etmeye yönelik basit doğrusal regresyon modelinin geliştirilmesi ve validasyonunu kapsamıştır. Çalışma boyunca elde edilen ana bulgular ve değerlendirmeler aşağıda özetlenmiştir:

Ana Bulgu ve İş Perspektifi:

1. *Güçlü Negatif İlişki:*
   weight ile mpg arasında başlangıçta tespit edilen -0.84%lük güçlü negatif korelasyon, araç ağırlığının yakıt tüketimi üzerinde dominant bir etkiye sahip olduğunu açıkça göstermiştir. Bu, otomotiv sektöründe tasarım ve üretim süreçlerinde ağırlık optimizasyonunun yakıt verimliliği açısından kritik bir faktör olduğunu vurgulamaktadır.

2. *Dönüşümlerin Önemi:*
   Orijinal verilerle kurulan modelin regresyon varsayımlarını (normallik ve homoskedastisite) ihlal etmesi, veri dönüşümlerinin gerekliliğini ortaya koymuştur. Sqrt(weight) ve Log(mpg) dönüşümlerinin uygulanmasıyla modelin R^2 değeri %69.9%dan %77.2%ye yükselmiş ve artık varsayımları sağlanmıştır. Bu, istatistiksel olarak geçerli ve daha açıklayıcı bir model elde edildiği anlamına gelmektedir. İş perspektifinden bakıldığında, bu tür dönüşümler, daha doğru tahminler ve dolayısıyla daha bilinçli iş kararları alınmasına olanak tanır.

3. *İyileşmiş Tahmin Performansı:*
   Dönüştürülmüş modelin eğitim setindeki MAE değeri 3.016, test setindeki MAE değeri ise 3.088 olarak gerçekleşmiştir. Orijinal modelin MAE değeri 3.230 iken, bu iyileşme, modelin hem eğitim hem de test verileri üzerinde daha düşük hata oranlarıyla tahmin yapabildiğini göstermektedir. Bu, yakıt tüketimi tahminlerinin daha güvenilir olduğu ve

örneğin yeni araç modellerinin yakıt verimliliği performansını öngörmede daha isabetli sonuçlar verebileceği anlamına gelir.

Model Sınırlılıkları ve Öneriler:

1. *Basit Doğrusal Model:*

   Bu çalışma basit doğrusal regresyon modeline odaklanmıştır. Gerçek dünyadaki yakıt tüketimi, sadece ağırlık gibi tek bir değişkene bağlı olmayıp, silindir sayısı, motor hacmi, beygir gücü, hızlanma, model yılı gibi birçok faktörden etkilenmektedir. Bu nedenle, modelin açıklayıcılık gücü ve tahmin doğruluğu, çoklu doğrusal regresyon veya daha karmaşık makine öğrenimi modelleri (örneğin, karar ağaçları, random forest, gradyan artırma modelleri) kullanılarak artırılabilir.

2. *Veri Seti Kapsamı:*

   Auto MPG veri seti, belirli bir zaman dilimindeki (1970%ler-1980%ler) araçları içermektedir. Günümüz araç teknolojileri ve yakıt verimliliği standartları önemli ölçüde farklılık göstermektedir. Modelin güncel araçlar için geçerliliğini artırmak amacıyla daha yeni ve kapsamlı veri setlerinin kullanılması önerilir.

Sonuç olarak, weight ve mpg arasındaki ilişkiyi modellemek için uygulanan veri dönüşümleri, modelin istatistiksel varsayımlarını sağlamada ve tahmin performansını iyileştirmede başarılı olmuştur. Bu proje, veri analizi ve modelleme süreçlerinde varsayım kontrolünün ve uygun veri dönüşümlerinin kritik rolünü vurgulayan değerli bir örnek teşkil etmektedir. Gelecekteki çalışmalar, daha karmaşık modeller ve güncel veri setleri ile bu analizi genişleterek daha kapsamlı ve pratik uygulamalar sunabilir.