MİMAR SİNAN GÜZEL SANATLAR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

KÜMELEME VE KARAR AĞACI: KANADA'DA SATIŞA ÇIKAN 2022 MODEL ARAÇLARI ÜZERİNE UYGULAMA

LİSANS TEZİ

Tarık KÜTÜK

Anabilim Dalı: İSTATİSTİK

Programı: LİSANS

Tez Danışmanı: Bilge BAŞER

HAZİRAN 2022

MİMAR SİNAN GÜZEL SANATLAR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

KÜMELEME VE KARAR AĞACI: KANADA'DA SATIŞA ÇIKAN 2022 MODEL ARAÇLARI ÜZERİNE UYGULAMA

LİSANS TEZİ

Tarık KÜTÜK

Anabilim Dalı: İSTATİSTİK

Programı: LİSANS

Tez Danışmanı: Bilge BAŞER

HAZİRAN 2022

Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Enstitüsü tez yazım kılavuzuna uygun olarak hazırladığım bu tez çalışmasında;

- Tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- Görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel etik kurallarına uygun olarak sunduğumu,
- Başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- Atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Ücret karşılığı başka kişilere yazdırmadığımı (dikte etme dışında), uygulamalarımı yaptırmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

ÖNSÖZ

Tüm hayatım boyunca gölgesini hiç üzerimden eksik etmeyen, ellerinden gelenin en iyisiyle beni bu yaşıma getiren çok sevgili aileme,

Okul hayatımın en başından itibaren arkadaşlığını esirgemeyen Seval'e,

Kadim dostum Abdulsamed'e,

Tez çalışmam boyunca öneri ve bilgi aktarımı konularında desteğini esirgemeyen değerli tez danışanım Dr. Bilge BAŞER'e

ve son olarak yaşanan her şeye rağmen devam etmenin bir yolunu bulan kendime,

Selam ve Teşekkür Ederim.

KÜMELEME VE KARAR AĞACI: KANADA'DA SATIŞA ÇIKAN 2022 MODEL ARAÇLARI ÜZERİNE UYGULAMA

ÖZET

Kümeleme, Sınıflama ve Tahminleme yöntemleri İstatistik alanında çok sık kullanılan yöntemlerdir. Bu araştırmanın amacı Kanada'da satışa çıkan 2022 model araçların özelliklerine göre kümelenmesi, araçların yakıt tipini tahmin eden bir karar ağacı modeli oluşturmak ve araçların ortalama yakıt tüketimini tahmin eden bir karar ağacı modeli oluşturmaktır.

Bu amaç doğrultusunda kullanılan veri seti önce Keşfedici Veri Analizi kapsamında incelenerek modellemeye hazır hale getirildikten sonra, veri setindeki araçlar kümelenecek; daha sonra ise yakıt tipine göre araçları sınıflayan ve ortalama yakıt tüketimini tahmin eden modeller kurulacaktır.

Anahtar Kelimeler; kümeleme, sınıflama, tahminleme, karar ağacı, k-means, hiyerarşik kümeleme, cart, keşfedici veri analizi, KVA

CLUSTERING AND DECISION TREE: IMPLEMENTING WITH 2022 MODEL VEHICLES SOLD IN CANADA

ABSTRACT

Clustering, Classification and Estimation methods are the most frequently used methods in the field of statistics. The purpose of this research is to cluster the 2022 model vehicles on sale in Canada according to their characteristics, build a decision tree model that predicts the fuel type of vehicles and build a decision tree model that predicts the average fuel consumption of vehicles.

After the data set used for this purpose is first examined within the scope of Exploratory Data Analysis and made ready for modeling, the cars in the data set will be clustered; Then, models that classify cars according to fuel type and estimate average fuel consumption will be established.

Keywords: Clustering, classification, prediction, decision tree, k-means, hierarchical clustering, cart, exploratory data analysis, EDA

İÇİNDEKİLER

ÇİZELGE LİSTESİ

ŞEKİL LİSTESİ

1. GİRİŞ

1.1 Veri Yapısı ve Hikayesi

Çalışma kapsamında kullanılan veri Kanada'da 2022 yılında satışa sunulan arabalara ait belirli özellikleri içermektedir. Bu özellikler;

- Arabaların markaları
- Arabaların modelleri
- Arabaların kasa tipi
- Arabaların motor hacimleri
- Arabalarının motorundaki silindir sayısı
- Arabaların şanzıman tipi ve ileri vites sayısı
- Arabaların yakıt tipi
- Arabaların şehir içi yakıt tüketimi (L/100 km)
- Arabaların şehir dışı yakıt tüketimi (L/100 km)
- Arabaların ağırlıklı (%55 şehir içi, %45 şehir dışı) ortalama yakıt tüketimi (L/100 km)
- Arabaların ağırlıklı (%55 şehir içi, %45 şehir dışı) ortalama yakıt tüketimi (MPG)
- Arabaların karbondioksit emisyon değerleri
- Arabaların karbondioksit değerlendirme puanları
- Arabaların egzoz dumanı değerlendirme puanları

1.2 Veri temizleme

Veride kasa tipi “two-seatter” olan arabaların performans araçları oldukları tespit edilmiştir. Performans araçlarında yakıt tüketimi önemsenmediği için bu kasa tipindeki araçlar veriden çıkarılmıştır.

Verideki “Transmission” değişkeninde arabaların şanzıman tipi ve vites sayıları birlikte bulunmaktadır. Bu değişkendeki bu iki bilgi ayrılmış ve “Transmission Type” isminde şanzıman tipine yönelik bir değişken oluşturulmuştur.

Verideki araçlarda Dizel ve Elektrikli araçların sayısı, Benzinli ve Hibrit araçlara göre oldukça az olduğu görülmüştür. Kullanılacak modellerde yanlılığa sebep olmaması için Dizel ve Elektrikli araçlar da veriden çıkarılmıştır.

1.3 Çalışmanın Amacı

Çalışma kapsamında;

- Verideki araçlar belirlenen özelliklerine göre kümelenmesi
- Araçları yakıt tipi benzinli veya hibrit olmasına göre sınıflanması
- Araçların ortalama yakıt tüketiminin tahminlenmesi amaçlanmaktadır.

2. KEŞFEDİCİ VERİ ANALİZİ

2.1 Nicel Değişkenlerin İstatistikleri

Nicel değişkenlere ait ortalama ve ortanca değerleri verilmiştir.

Çizelge 2.1 Nicel değişkenlerin ortalama ve ortancaları

Değişken adı	Ortalama	Ortanca
Motor Hacmi (EngineSize)	3.066	2.9
Şehir İçi Yakıt Tüketimi (Fuel Cons. City (L/100 km))	122.1	123
Şehir Dışı Yakıt Tüketimi (Fuel Cons. Highway (L/100 km))	92.03	92
Ortalama Yakıt Tüketimi (Fuel Consumption (Comb (L/100 km)))	108.5	108
Ortalama Yakıt Tüketimi (Fuel Consumption (Comb(MPG))	27.65	26

Verideki araçların motor hacimlerinin ortalaması 3.066 ve ortancası 2.9'dur. Araçların şehir içi yakıt tüketimlerinin ortalaması 122.1 ve ortancası 123'tür. Araçların şehir dışı yakıt tüketimi 92.03 ve ortancası 92'dir. Araçların ortalama yakıt tüketimi ortalamaları 108.5 ve ortancası da 108'dir. Araçların bir galonla gittikleri ortalama mil ise 27.65 ve ortancası 26'dır.

2.2 Nitel Değişkenlerin Frekans ve Yüzdelikleri

Nitel değişkenlerin sınıfları, sınıfların frekansları ve yüzdelikleri verilmiştir.

Çizelge 2.2 Araçların kasa tiplerine göre frekans ve yüzdelikleri

Kasa Tipi	Frekans	Yüzdelik
SUV: Small	141	22.89
Mid-size	90	14.61
SUV: Standard	87	14.12
Pickup truck: Standard	82	13.31
Subcompact	50	8.12
Full-size	44	7.14
Compact	39	6.33
Minicompact	39	6.33
Pickup truck: Small	16	2.60
Station wagon: Small	10	1.62
Special purpose vehicle	8	1.30
Minivan	5	0.81
Station wagon: Mid-size	5	0.81

Çizelge 2.3 Araçların yakıt tiplerine göre frekans ve yüzdelikleri

Yakıt Tipi	Frekans	Yüzdelik
Benzin	306	51.95
Hibrit	283	48.05

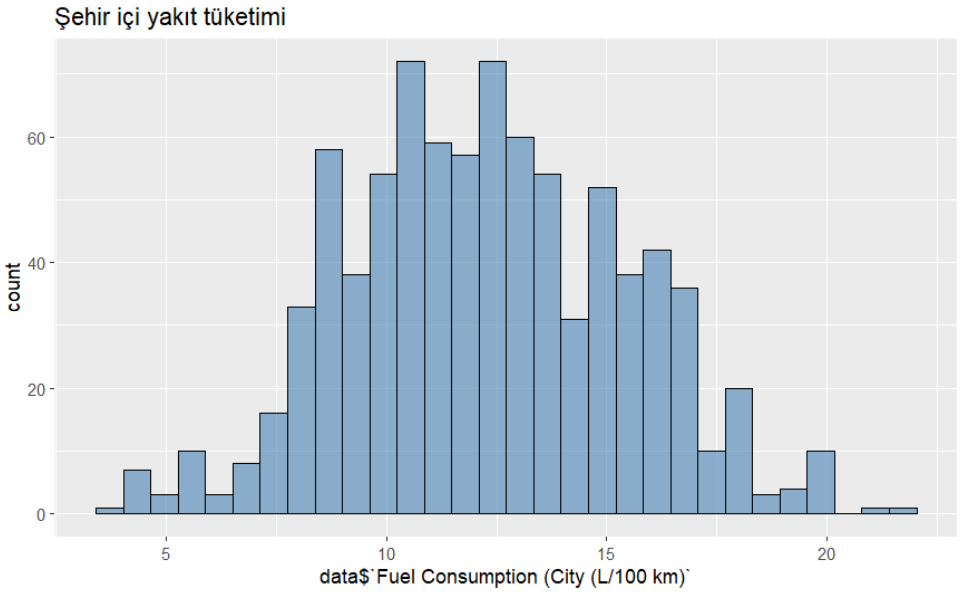
Çizelge 2.4 Şanzıman tiplerine göre frekans ve yüzdelikler

Şanzıman Tipi	Frekans	Yüzdelik
Sıralı (AS)	345	40.45
Otomatik (A)	227	26.61
Yarı Otomatik (AM)	106	12.43
CVT	96	11.25
Düz (M)	79	9.26

2.3 Nicel Değişkenlerin Dağılımı

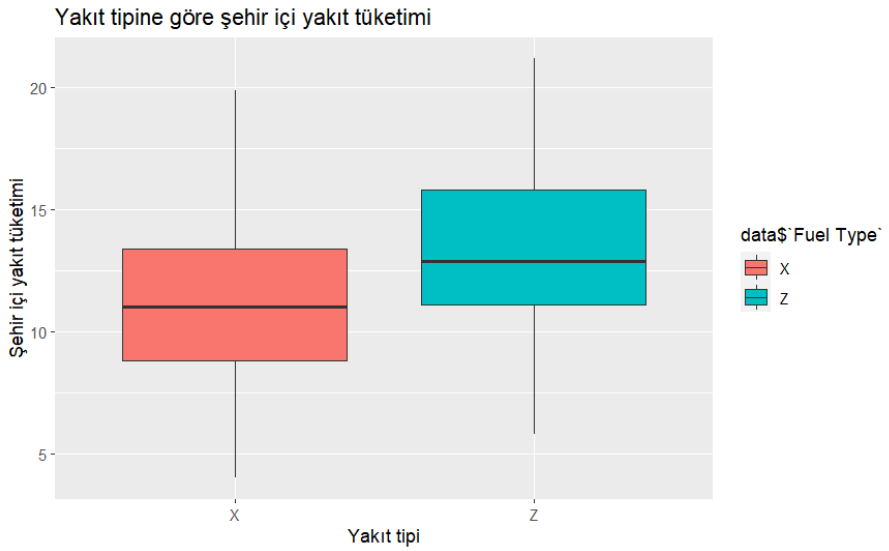
Yakıt tüketimine dair olan nicel değişkenlerin dağılımları incelenecektir.

2.3.1 Şehir İçi Yakıt Tüketimi



Şekil 2.1 Şehir İçi Yakıt Tüketimi Histogramı

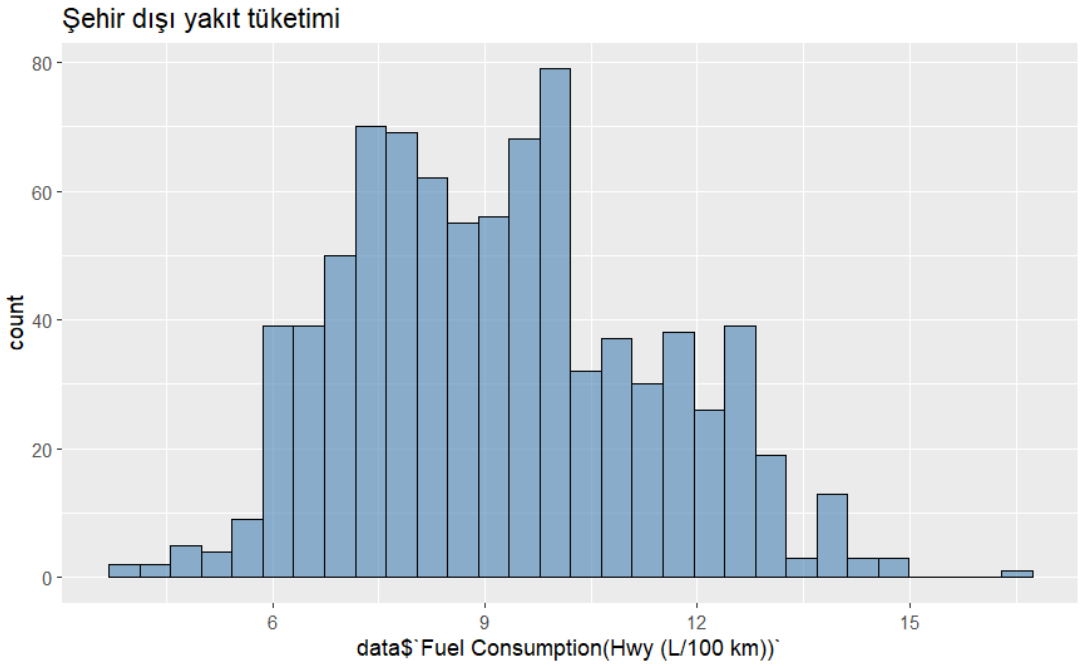
Şehir içi yakıt tüketimi histogramı verilmiştir. Değişkenin yaklaşık normal dağıldığı söylenebilir. Araçların şehir içi yakıt tüketiminin 10 ve 12,5 litre civarında bir yoğunluğu olduğu söylenebilir.



Şekil 2.2: Yakıt tipine göre şehir içi yakıt tüketimi kutu grafiği

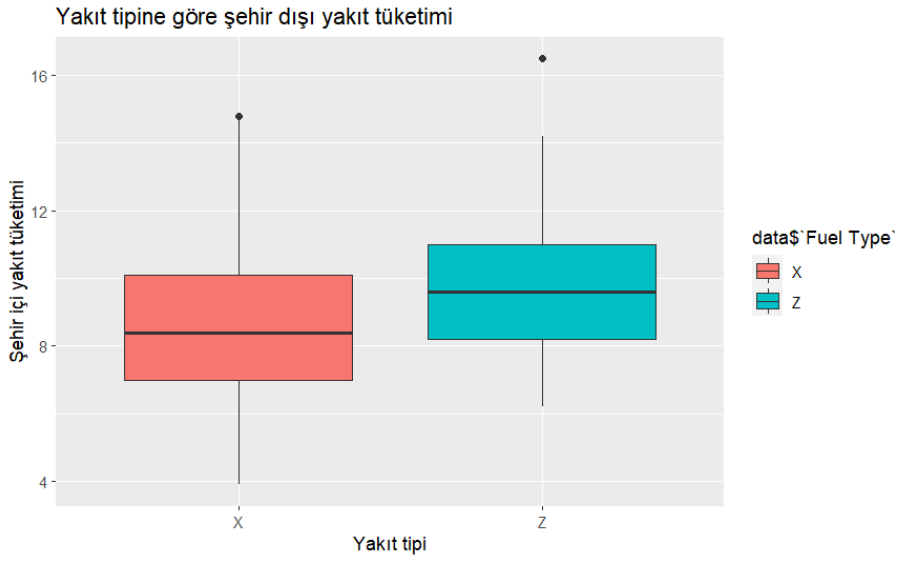
Yakıt tiplerine göre ("x" hibrit, "z" benzin) şehir içi yakıt tüketimi kutu çizimi verilmiştir. Benzinli araçların hibritlere göre yakıt tüketiminin daha yüksek olduğu yorumu yapılabilir. Bunun yanında iki grubun da yayılımının yakın olduğu söylenebilir. Benzinli grubun sağa çarpık bir dağılımı olduğunu, araçların daha çok ortalamanın altında yakıt tüketimine sahip olduğu söylenebilir.

2.3.2 Şehir Dışı Yakıt Tüketimi



Şekil 2.3 Şehir dışı yakıt tüketimi histogramı

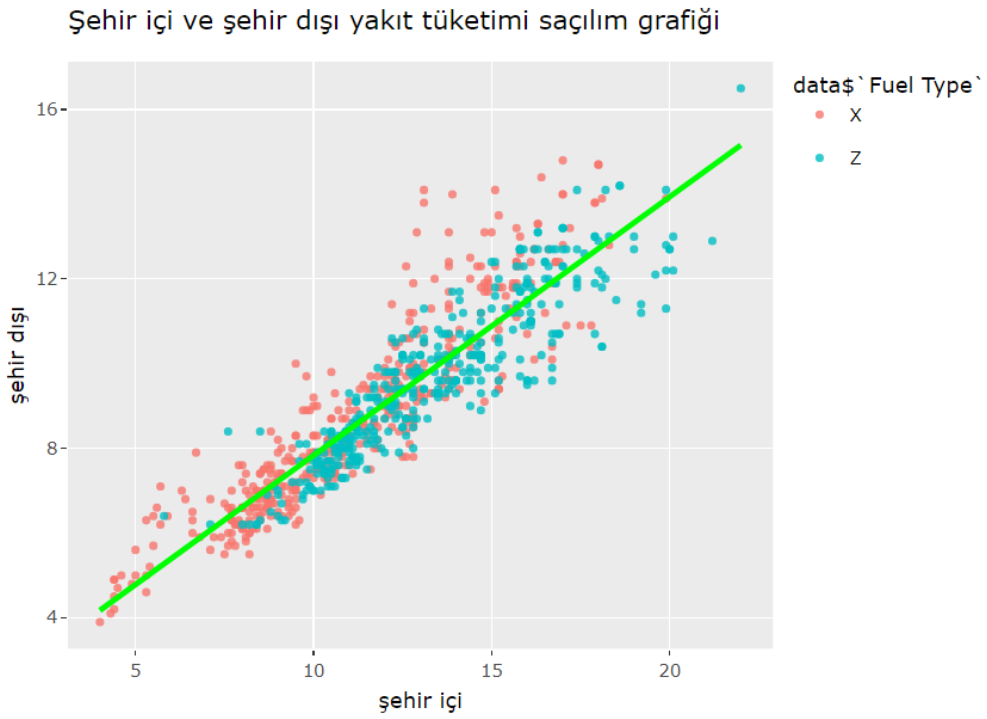
Şehir dışı yakıt tüketimi histogramı verilmiştir. Araçların 7,5 litre ile 10 litre arasında bir yoğunluğu olduğu söylenebilir. Sağa çarpık bir dağılım olduğu görülmüştür. Uç değer varlığından söz edilebilir.



Şekil 2.4 Yakıt tiplerine göre şehir dışı yakıt tüketimi

Yakıt tiplerine göre şehir dışı yakıt tüketimi kutu çizimleri verilmiştir. Şehir dışı değerlerinde de benzinli araçların hibritlere göre yakıt tüketiminin daha fazla olduğu görülmüştür. Fakat bu sefer iki grubun da sağa çarpık olduğu, benzinli grubun ise yayılımının, hibrit gruba göre oldukça az olduğu görülmüştür. İki grup arasındaki fark ise şehir dışı yakıt tüketimine göre azalmıştır. Bunun yanında iki grubun da yakıt tüketimi şehir içine göre azalmıştır. İki grupta da birer tane aykırı değer burada da göze çarpmaktadır.

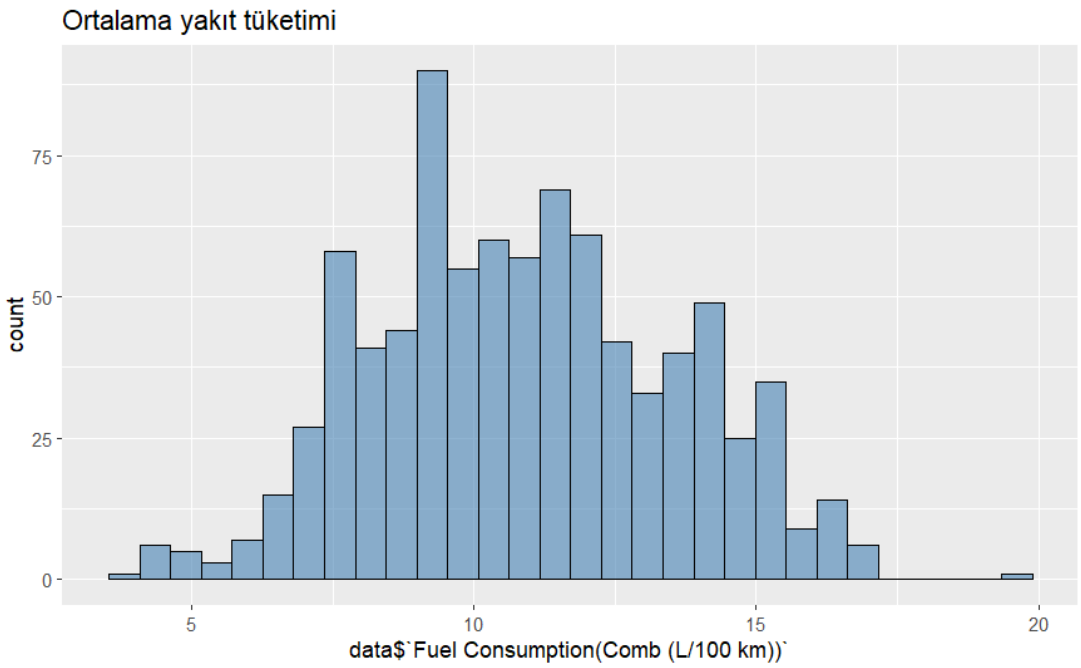
2.3.3 Şehir İçi Yakıt Tüketimi ile Şehir Dışı Yakıt Tüketimi Arasındaki İlişki



Şekil 2.5 Şehir içi ve şehir dışı yakıt tüketimleri saçılım grafiği

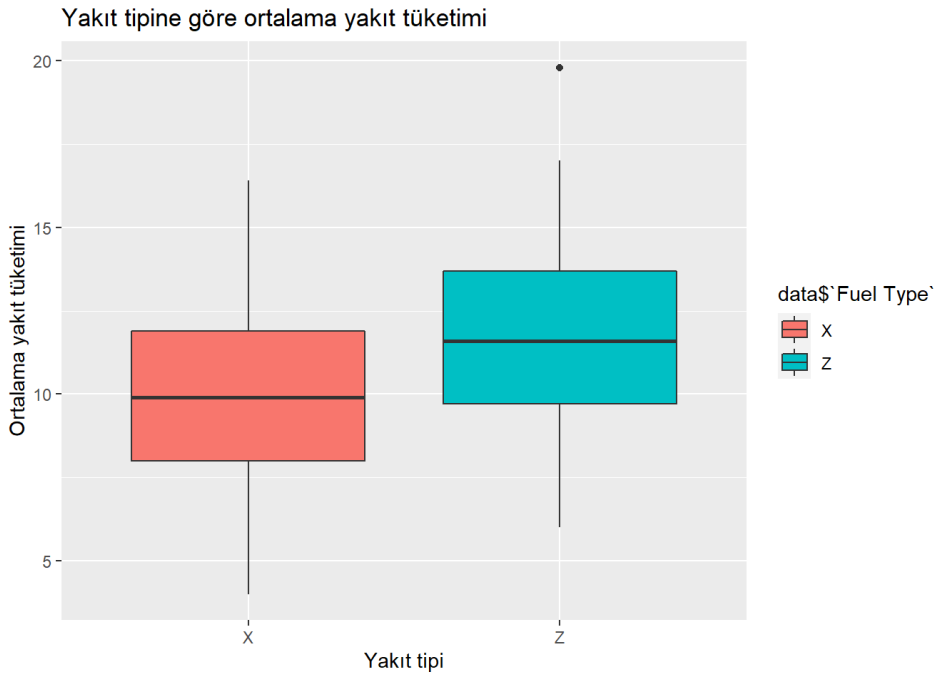
Şehir içi ve şehir dışı yakıt tüketimi saçılım grafiği verilmiştir. Değişkenler arasındaki doğrusal ilişki ile değişen varyans göze çarpmaktadır. Bunun yanında şehir içi yakıt tüketimi 22 lt, şehir dışı yakıt tüketimi 16.5 lt olan bir gözlemin burada da uç değer olduğu görülmektedir.

2.3.4 Ortalama Yakıt Tüketimi



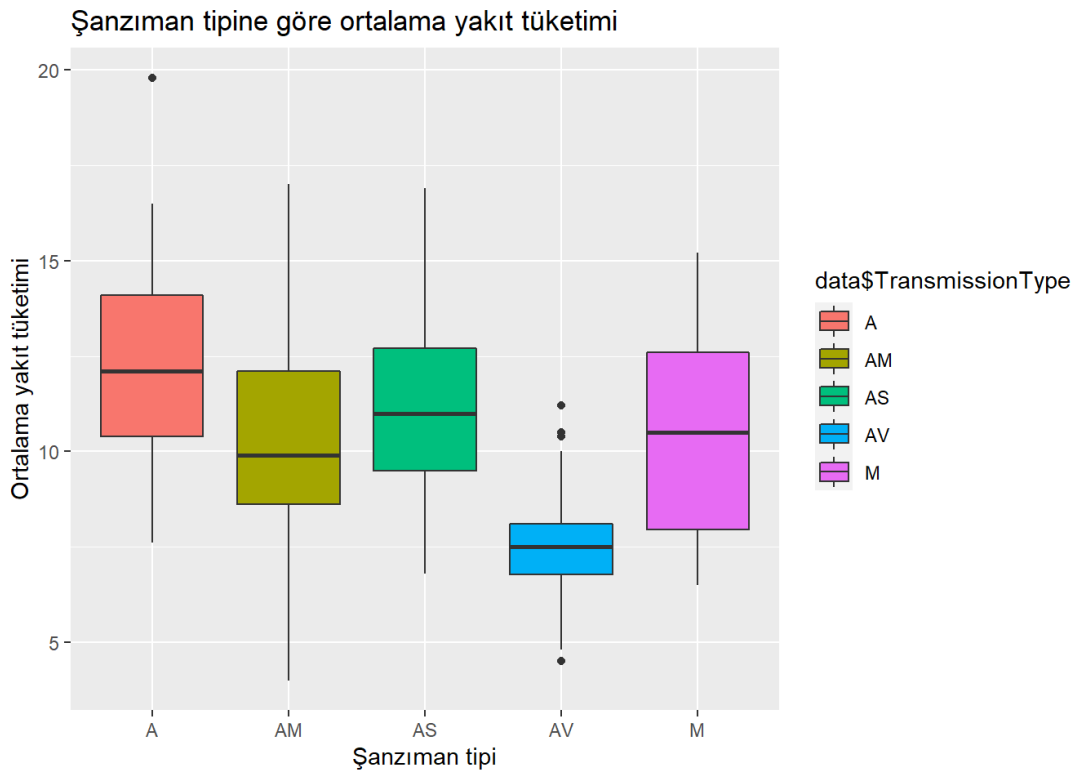
Şekil 2.6 Ortalama yakıt tüketimi histogramı

Ortalama yakıt tüketimi histogramı verilmiştir. Araçların ortalama yakıt tüketiminin 7.5-12.5 lt civarında bir yoğunluğa sahip olduğu söylenebilir. Yaklaşık normal dağılım gösterdiği söylenebilir. Uç değer burada da göze çarpmaktadır.



Şekil 2.7 Yakıt tiplerine göre ortalama yakıt tüketimi

Beklenildiği üzere ortalama yakıt tüketiminde de benzinli araçların hibritlere göre daha yüksek yakıt tüketimine sahip olduğu anlaşılmaktadır. Hibrit grubun ortalama yakıt tüketimi yayılımı, benzinli gruba göre daha fazla olduğu görülmüştür. Benzinli grubun, hibrit gruba göre daha fazla sağa çarpık olduğu ve bu gruptaki aykırı gözlemin varlığının devam ettiği görülmüştür.



Şekil 2.8 Şanzıman tiplerine göre ortalama yakıt tüketimi

Otomatik şanzımana sahip araçların ise yakıt tüketimi bakımından en yüksek değerlere sahip olduğu ve yaklaşık normal dağılım gösterdiği yani gözlemlerin grup içerisinde 12 lt yoğunlaştıklarını fakat yaklaşık 17 lt civarına kadar dağılım kuyruğunun ilerlediği anlaşılır. 19 lt'ye yakın bir uç değerin olduğu da görülmektedir.

Yarı otomatik grubun ise sağa çarpık bir dağılım gösterdiğini ve 10 lt civarında yoğunlaştığını söyleyebiliriz. Bununla birlikte dağılım kuyruklarının da uzun olduğunu yani bu şanzımana sahip araçların ortalama yakıt tüketimlerinin oldukça değişken olduğu söylenebilir.

Sıralı şanzıman grubunun ise sağa çarpık bir dağılım gösterdiğini ve otomatik gruptan sonra en fazla yakıt tüketimine sahip grup olduğu anlaşılmaktadır.

CVT şanzımana sahip olan araçların ise en düşük yakıt tüketimine sahip olduğu 4 adet uç gözlemi olduğu ve yaklaşık normal dağıldığı görülmüştür. DAG değeri en küçük olan grup olduğu görülmüştür ki bu en az değişkenlik gösteren grup olduğu anlamına

gelir. Buradan CVT şanzımanın diğer şanzıman tiplerine göre yakıt tüketiminde daha stabil değerler aldığı anlamına gelir.

Manuel şanzımana sahip olan araçların ise en fazla ortalama yakıt tüketimine sahip 3. grup olduğu ve ortanca çevresinde bir yoğunlaşma olduğu görülmüştür. Diğer gruplara göre DAG değeri en yüksek olan gruptur ki bu en fazla değişkenlik gösteren grup olduğu anlamına gelir.

2.4 Uç Değerler

Daha önce incelenen histogram, kutu çizimleri ve saçılım grafiklerinde uç değer olduğu görülen gözlemler incelenecektir.

Çizelge 2.5 Uç değer olduğu tespit edilen gözlemler

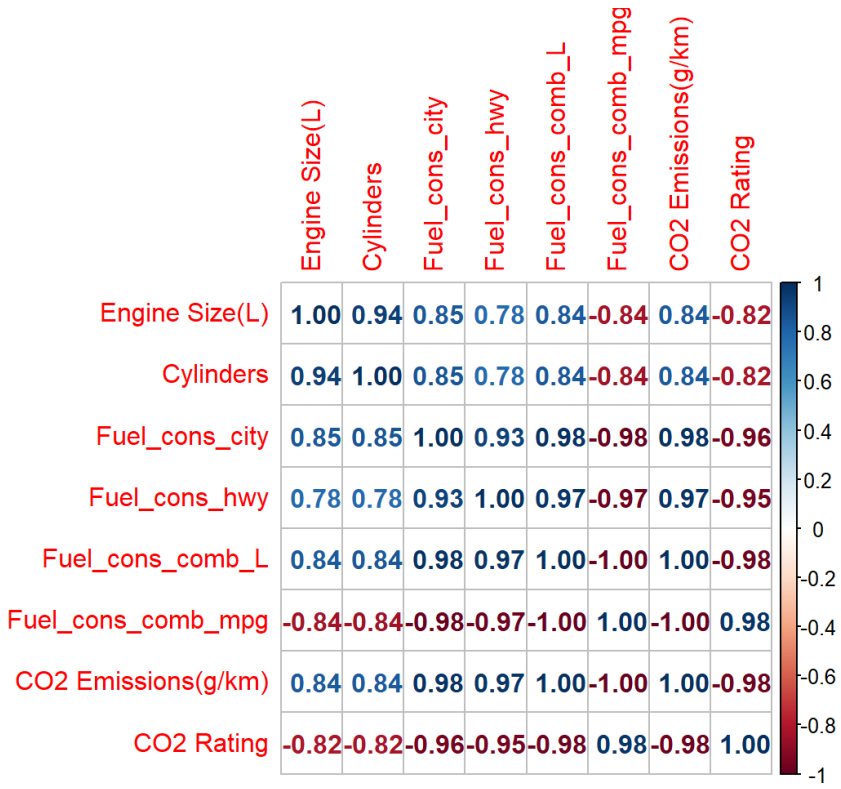
Marka	Model	Ortalama yakıt tüketimi	Şehir dışı yakıt tüketimi
Subaru	WRX AWD	11,2	9,4
Subaru	Ascent AWD	10,5	9,0
Nissan	Murano AWD	10,4	8,5
Ram	1500 4X4 TRX	19,8	16,5
Ford	F-150 Raptor	15,8	14,8
Toyota	Prius	4,5	4,7

Uç değerlere sahip gözlemler incelendiğinde burada yapılacak analizler için uç değer sayılabilecek tek gözlemin Ram markasının 1500 4X4 TRX aracı olduğu görülmüştür. Yapılan araştırmada bu aracın performansa yönelik bir araç olduğu görülmüştür. Performans araçlarında yakıt tüketimi önemli görülmediği için bu araç da çalışma kapsamından çıkarılacaktır.

Ram TRX haricindeki uç değerli araçların yakıt tüketimlerinde verideki diğer araçlara göre bir uçluk söz konusu olmadığından bu araçlarla ilgili bir işlem yapılmayacak, bu araçlar çalışma kapsamında kalacaklardır.

2.5 Korelasyon

Verideki nicel değişkenler normal dağılmadığı için Spearman Korelasyon Katsayısı kullanılmıştır. Korelasyon matrisi şu şekildedir.



Şekil 2.9 Spearman Korelasyon Katsayısı matrisi

Korelasyon matrisi incelendiğinde değişkenler arasında yüksek korelasyon olduğu görülmüştür.

Araçların motor hacimleri arttıkça silindir sayısı, şehir içi yakıt tüketimi, şehir dışı yakıt tüketimi, ortalama yakıt tüketimi (L) ve karbondioksit emisyonu artarken, ortalama yakıt tüketimi (mpg) ve karbondioksit değerlendirme puanın azalmaktadır.

2.6 Kodlama

Kullanılacak yöntemlerin doğru çalışabilmesi için kullanılacak nitel değişkenlerin düzeyleri kodlanmıştır.

Çizelge 2.6: Nitel değişkenlerin kodlaması

Değişken adı	Değişken düzeyi	Kodlaması
Vehicle Class (Kasa tipi)	Compact	0
	Full-size	1
	Mid-size	2
	Minicompact	3
	Minivan	4
	Pickup-Truck: Small	5
	Pickup-Truck:Standart	6
	Special Purpose Vehicle	7
	Station Wagon: Mid-size	8
	Station Wagon: Small	9
	Subcompact	10
	SUV: Small	11
	SUV: Standart	12
Fuel Type	Hibrit (X)	1
	Benzin (Z)	2
Transmission Type	Otomatik (A)	0
	Yarı Otomatik (AM)	1
	Sıralı (AS)	2
	CVT (AV)	3
	Düz (M)	4

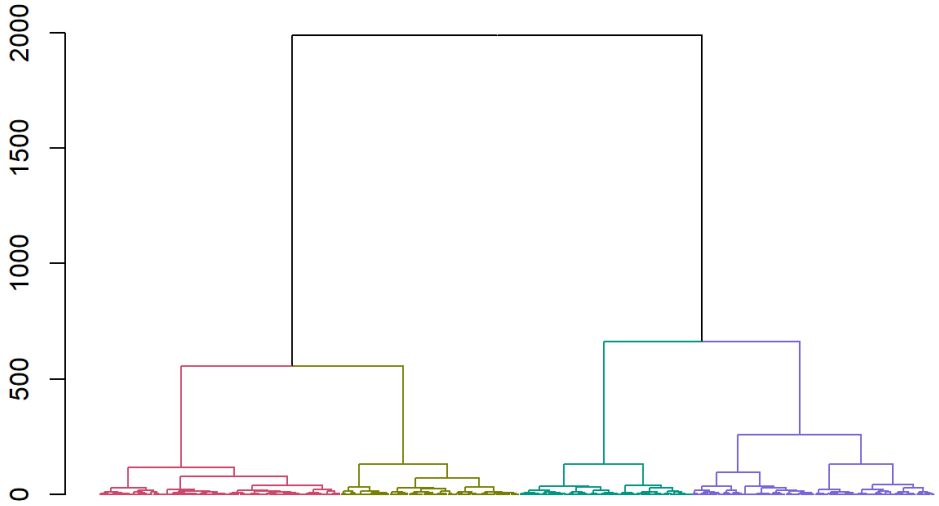
3. KÜMELEME

Verideki araçların sahip oldukları özelliklere göre kümelenmesi amaçlanmaktadır. Kümelemede kullanılacak değişkenler kasa tipi, motor hacmi, yakıt tipi, ortalama yakıt tüketimi (L), karbondioksit değerlendirme puanı ve şanzıman tipidir.

3.1 Hiyerarşik Kümeleme

Araçların kaç kümeye ayrılacağına karar vermek için önce Ward methoduyla Hiyerarşik Kümeleme yapılmıştır.[1]

Hiyerarşik Kümeleme sonucunda ulaşılan dendogram şu şekildedir.



Şekil 3.1: Hiyerarşik Kümeleme dendogramı

Dendograma bakarak araçların 4 kümeye ayrılmasına karar verilmiştir.

3.2 K-Means Kümeleme

Araçların 4 kümeye ayrılmasına karar verilmiştir.[2] K-Means Kümeleme modeli kurulmuş ve gözlemler benzer özelliklerine göre 4 kümeye ayrılmıştır. Küme merkezlerinin değişken değerleri şu şekildedir.

Çizelge 3.1: K-Means Yöntemi küme merkezleri

Değişkenler/kümeler	1. küme	2. küme	3. küme	4. küme
Kasa Tipi	1.25	10.57	11.42	3.77
Motor Hacmi	2.22	2.20	4.15	4.13
Yakıt Tipi	1.57	1.61	1.41	1.44
Ortalama Yakıt Tüketimi (L)	8.30	9.20	13.32	12.96
Karbondioksit Değerlendirme Puanı	6.12	5.46	3.30	3.47
Şanzıman Tipi	2.28	1.73	1.44	1.83

Araçların en genelde ortalama yakıt tüketimi fazla ve az olanlar olarak 2'ye ayrılıp, daha sonra da diğer özelliklerine göre bu 2 kümenin de 2'ye ayrıldığı söylenebilir. 1. ve 2. kümede yakıt tüketimi düşük araçların, 3. ve 4. kümede ise yakıt tüketimi yüksek araçların kümelendiği görülmüştür.

3.2.1 Nicel Değişkenler Bakımından Küme Merkezlerinin Yorumu

Ortalama yakıt tüketimi en yüksek küme 3. küme olmuştur. Kümedeki araçların ortalama yakıt tüketimlerinin ortalamasının 13,33 lt olduğu görülmüştür. Bu kümedeki araçların motor hacimleri ortalama 4,14 lt iken karbondioksit değerlendirme puanı ise 3,30 olduğu görülmüştür.

Ortalama yakıt tüketimi yüksek olan diğer küme ise 4. kümedir. Bu kümede ise araçların yakıt tüketiminin ortalama 12.96 lt olduğu görülmüştür. Bu kümedeki araçların motor hacimleri ortalama 4.13 lt ve karbondioksit değerlendirme puanı ortalaması 3.47 ile değerlendirmesi en kötü olan araçların olduğu kümedir.

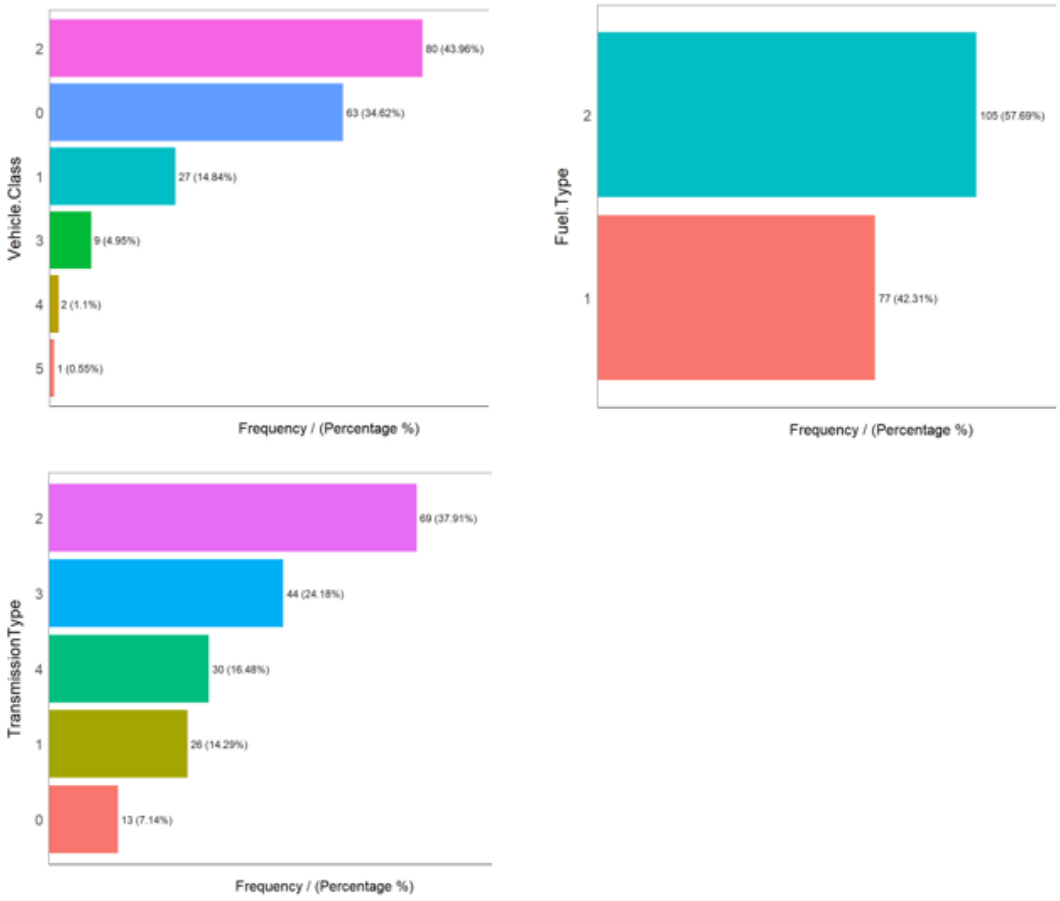
2. küme ise yakıt tüketimi düşük kümelerin fazla yakan kümesi olmuştur. Bu kümedeki araçların yakıt tüketim ortalaması 9,20 lt olduğu görülmüştür. Küme, hacim bakımından en küçük motorlu araçları içermektedir. Araçların ortalama motor hacmi

2,20 lt'dir. Karbondioksit değerlendirme puanı ise 5,46 ile bu değişkene göre 2. en iyi kümedir.

1. küme ise yakıt tüketimi en düşük araçların olduğu küme olmuştur. Bu kümedeki araçların yakıt tüketimi ortalaması 8.30 lt'dir. Bu değer ile en az yakan küme olduğu görülmüştür. Kümedeki araçların motor hacimleri ortalama 2.22 lt'dir. Karbondioksit değerlendirme puanı 6.12'dir ki bu karbondioksit salınımı bakımından en iyi küme olduğunu gösterir.

3.2.2 Nitel Değişkenler Bakımından Küme Frekanslarının Yorumu

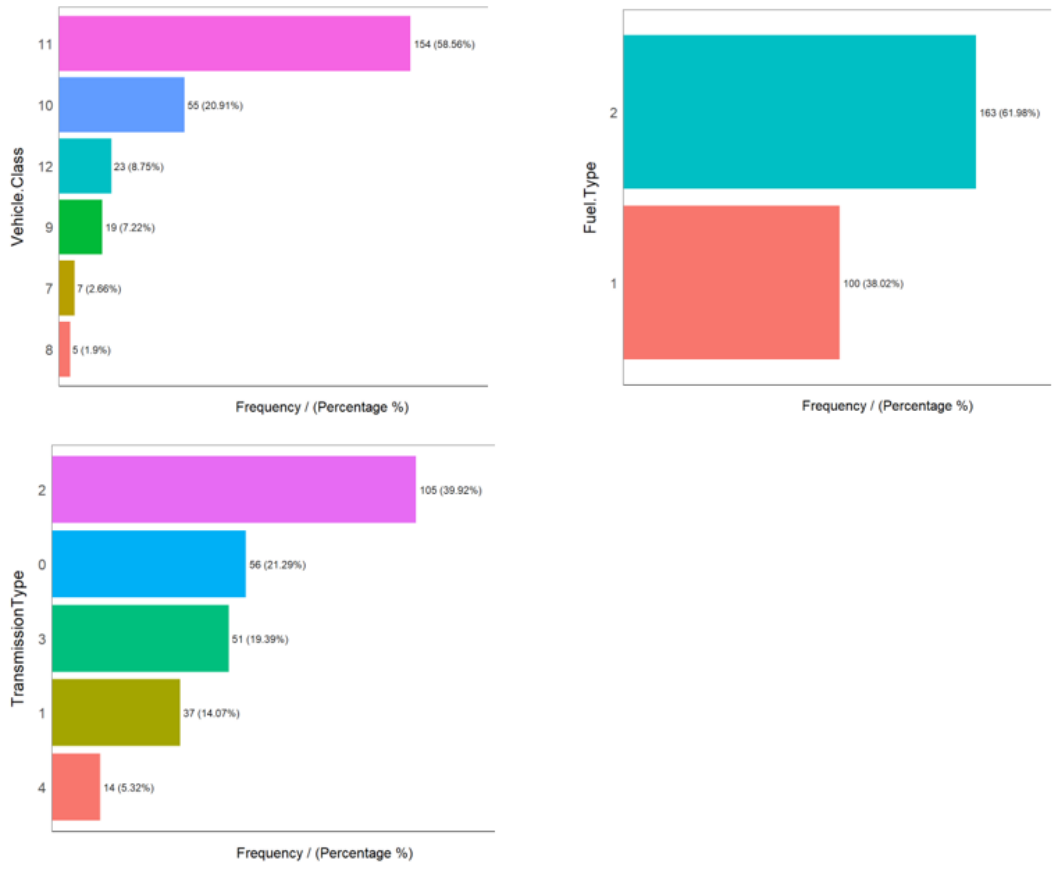
1. Kümedeki araçların nitel değişken frekansları verilmiştir



Şekil 3.2: 1. Kümedeki araçların nitel değişken frekansları

1. kümedeki araçların %43'ünün geniş hacimli sedan tipteki araçlar ve %34'ünün de küçük hacimli sedan tipteki araçlar olduğu görülmüştür. Bu araçların %57'sinin benzinli, %47'sinin ise hibrittir. Aynı zamanda araçların %37'sinin şanzımanı sıralı iken, %24' CVT olduğu görülmüştür.

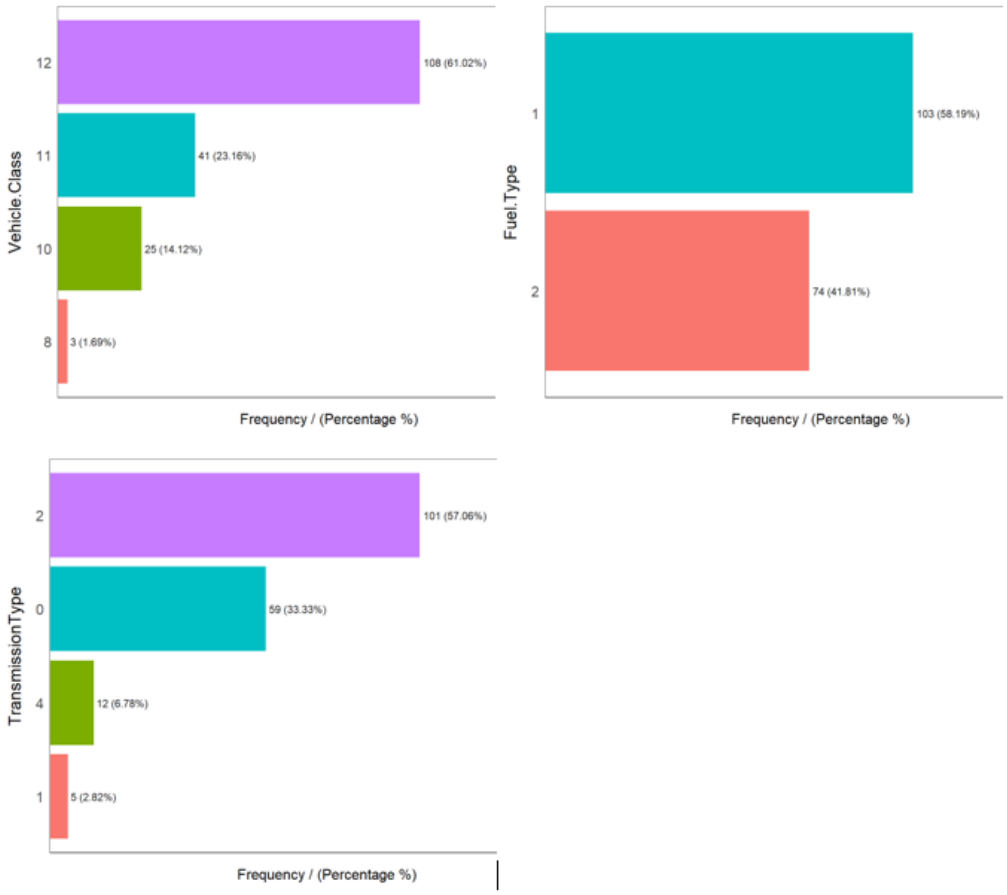
2. Kümedeki araçların nitel değişken frekansları verilmiştir



Şekil 3.3: 2. Kümedeki araçların nitel değişken frekansları

2. kümedeki araçların %58'inin küçük SUV tipindeki araçlar ve %20'sinin orta hacimli sedan araçlar olduğu görülmüştür. Araçların %61'inin benzinli %38'inin ise hibrit olduğu görülmüştür. Araçların %39'u sıralı şanzımana sahipken, %21'i de otomatik şanzımanlıdır.

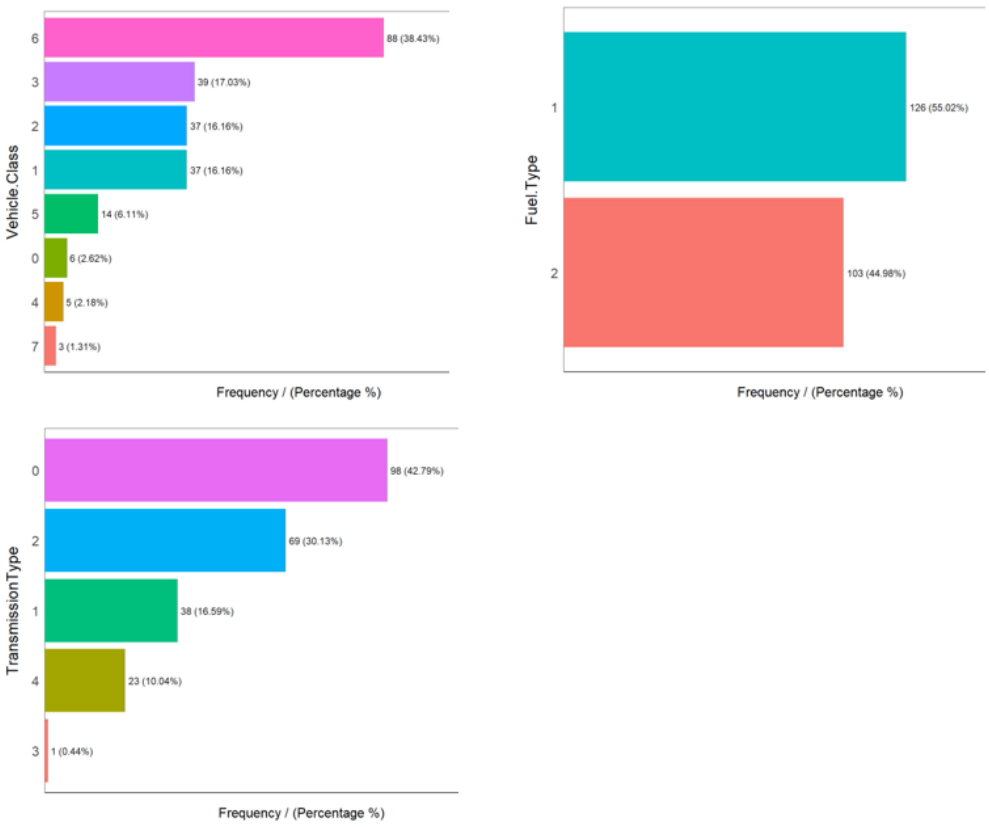
3. kümedeki araçların nitel değişken frekansları verilmiştir.



Şekil 3.4: 3. Kümedeki araçların nitel değişken frekansları

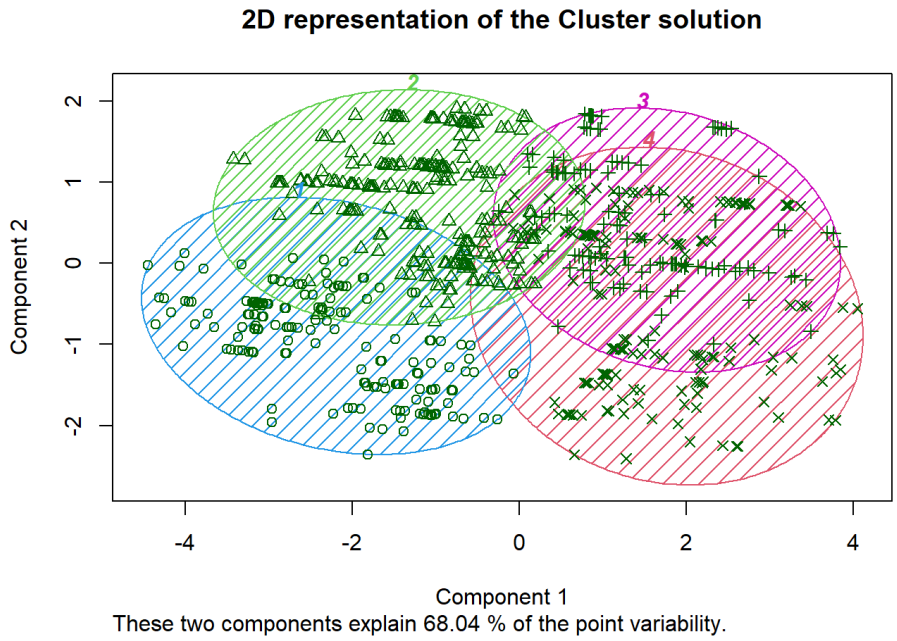
3. yani yakıt tüketimi en yüksek olan kümedeki araçların kategorik değişkenlerdeki frekansları verilmiştir. Kümedeki araçların %61'inin büyük SUV, %23'ünün ise küçük SUV olduğu görülmüştür. Araçların %58'i hibritken, %41'i benzinlidir ve araçların %57'si sıralı şanzımanlıyken, %33'ü otomatik şanzımanlıdır.

4. kümedeki araçların nitel değişken frekansları verilmiştir.



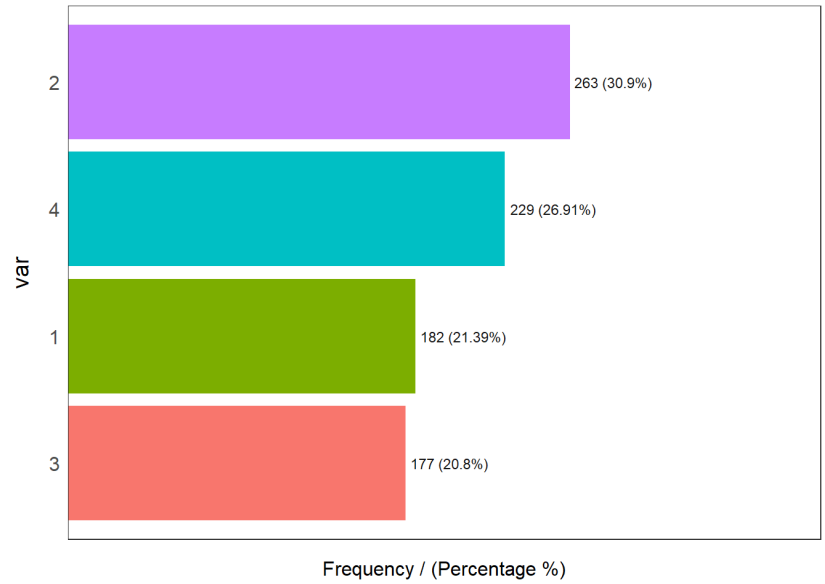
Şekil 3.5: 4. Kümedeki araçların nitel değişken frekansları

4. kümedeki araçlarınsa %38'inin büyük hacimli kamyonet oldukları görülmüştür. Kümedeki araçların %55'inin benzinli ve %45'inin de hibrit olduğu görülmüştür. Araçların %42'si otomatik ve %30'u da sıralı şanzımana sahiptir.



Şekil 3.6 Kümelerin 2 boyutlu saçılımı

Kümeleme yapılan verinin varyansını %68,04 açıklayan 2 bileşene göre kümeleme sonuçlarının saçılım grafiği verilmiştir. Burada 1. ve 3. kümenin birbirlerine göre ayrık iken diğer kümelerin birbirleriyle ortak özellikleri olduğu görülmüştür.



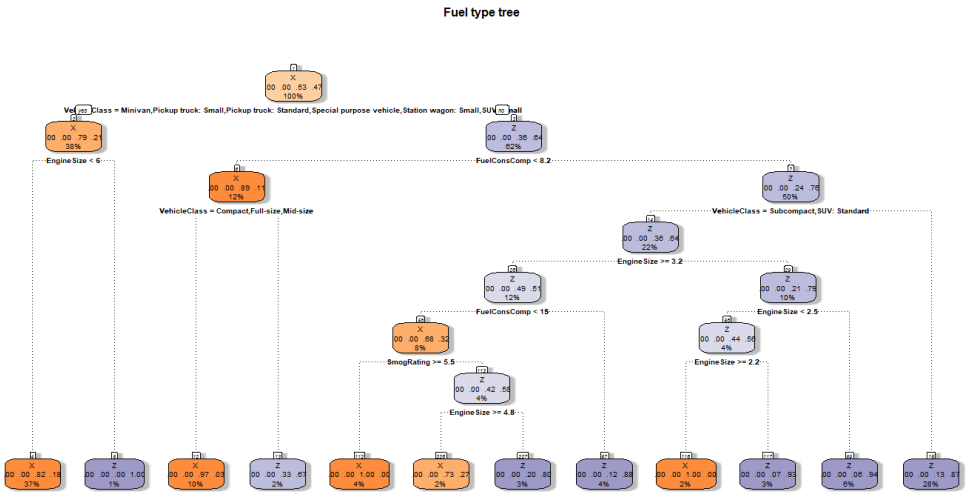
Şekil 1.7 Kümelerdeki gözlem sayıları

Kümelerdeki gözlem sayıları verilmiştir. En fazla araç 2. Kümedeyken en az aracın ise 3. Kümede olduğu görülmektedir.

4. Sınıflandırma

Verideki araçlar yakıt tipine göre sınıflandırılacaktır. Sınıflandırma için Karar Ağacı yöntemi kullanılmıştır. [3] Kurulacak sınıflandırma modelinin performansını ölçmek ve modeli denetleyebilmek için veri seti %70'i eğitim ve %30'u da test olmak üzere rastgele 2 parçaya bölünmüştür. Eğitim verisi ile model kurulmuştur ve test kümesiyle de model test edilecektir.

Yakıt tipine göre sınıflama yapan modelin kullandığı değişkenler kasa tipi, motor hacmi, silindir sayısı, ortalama yakıt tüketimi (L), karbondioksit emisyonu, karbondioksit değerlendirme puanı, egzoz dumanı değerlendirme puanı ve şanzıman tipidir.



Şekil 2.1: Yakıt tipini sınıflandıran karar ağacı modeli

Kurulan karar ağacı modeli verilmiştir. Test edilen modelin karmaşıklık matrisi şu şekildedir.

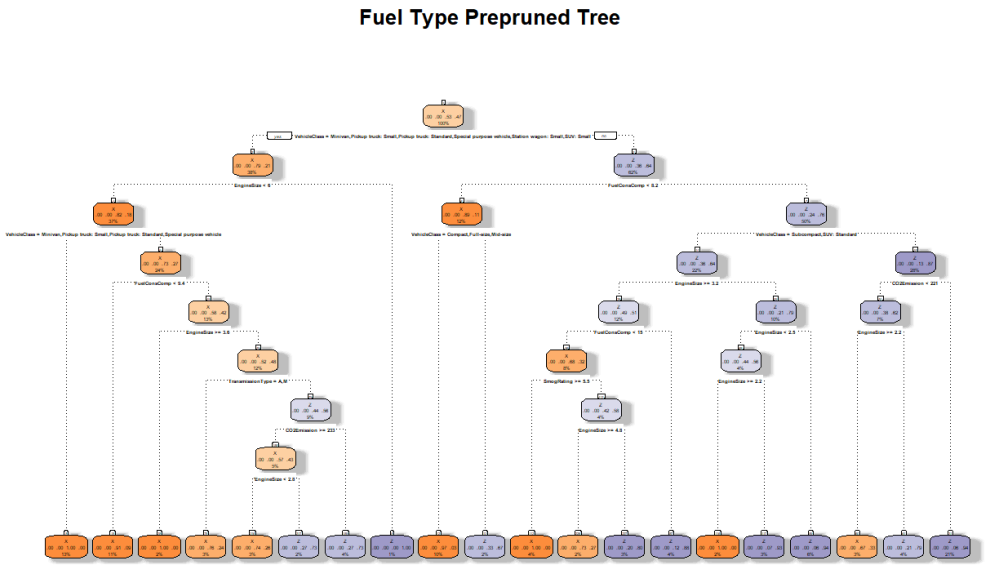
Çizelge 4.1 Test verisi ile test edilen modelin karmaşıklık matrisi

	Gerçekte hibrit olan	Gerçekte benzinli olan
Modelin hibrit atadığı	115	26
Modelin benzinli atadığı	16	98

Kurulan model daha önce görmediği test verisi ile çalıştırılmış ve %83 doğru atama yaptığı, yani test verisindeki 255 gözlemin 213'ünün yakıt tipini doğru sınıflamıştır.

4.1 Sınıflandırma Ağacının Budanması

Modelin doğru sınıflama yüzdesini yükseltmek için ağaç budanacaktır. [4] Budama yöntemi olarak önsel budama kullanılmıştır. Önsel budama yöntemi ile budanan karar ağacı modeli şu şekildedir.



Şekil 4.2: Önsel budama yapılan yakıt tipini sınıflandıran karar ağacı modeli

Önsel budama yapılmış karar ağacı modeli test verisi ile çalıştırılmıştır ve ulaşılan karmaşıklık matrisi şu şekildedir.

Çizelge 4.2 Test verisi ile test edilen budanan karar ağacı modeli

	Gerçekte hibrit olan	Gerçekte benzinli olan
Modelin hibrit atadığı	114	18
Modelin benzinli atadığı	17	106

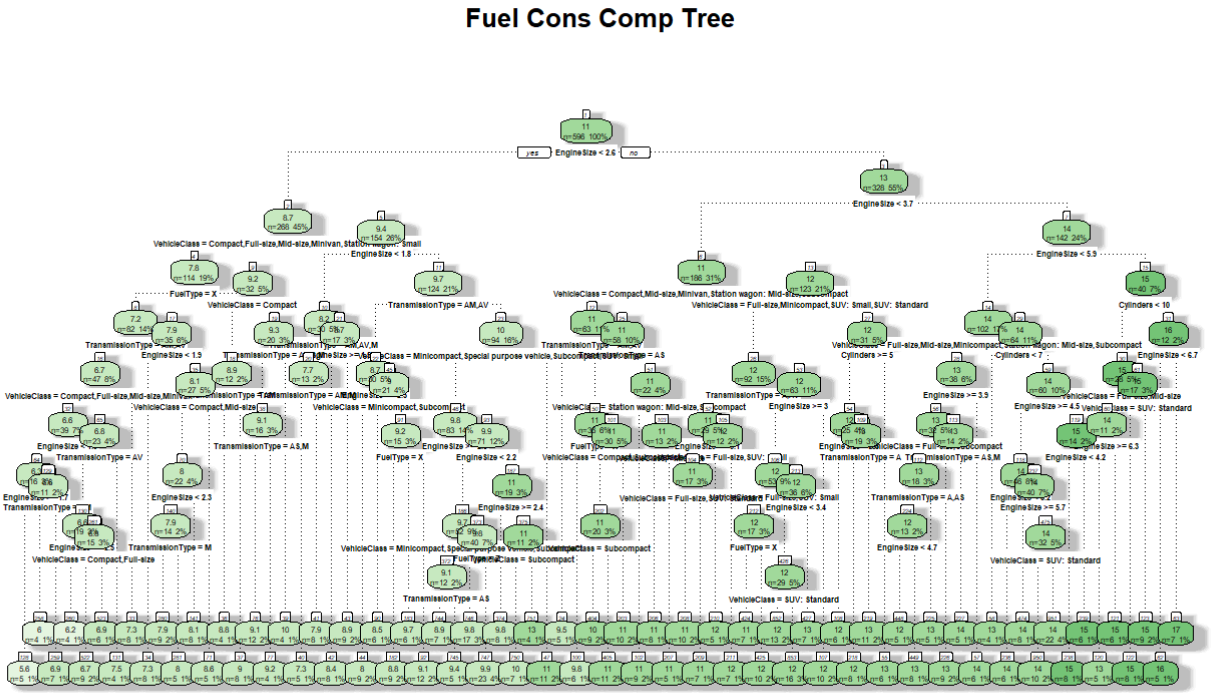
Önsel budama yapılan modelin doğru atama yüzdesinin %83'ten %86'ya çıktığı görülmüştür. Model test verisindeki 255 gözlemin 220'sini doğru sınıflandırmıştır.

Önsel ve Sonsal budama için farklı seçenekler denenmiş fakat doğruluk değerinin yükselmediği görülmüştür. Model, gerekli olan değişken değerleri verilen araçların yakıt tipini %86 doğru sınıflamaktadır.

5. TAHMİNLEME

Verideki araçların ortalama yakıt tüketimi (L) tahminlenecektir. Tahminleme için de Karar Ağacı yöntemi kullanılmıştır. [5] Daha önce veri bölünerek oluşturulan eğitim ve test veri kümeleri burada da kullanılacaktır.

Ortalama yakıt tüketimini tahminleyen model için kasa tipi, motor hacmi, silindir sayısı, yakıt tipi ve şanzıman tipi değişkenleri kullanılmıştır. Çalışmadaki amaçlardan biri de yeni bir araba yapıldığında ortalama yakıt tüketiminin ne olacağını tahmin etmek olduğu için bu modelde karbondioksit emisyonu gibi değişkenler kullanılmamıştır.



Şekil 4.3: Ortalama yakıt tipini tahminleyen karar ağacı modeli

Kurulan karar ağacı modeli verilmiştir. Bu model ile test verisi çalıştırılmış ve belirtme katsayısı, ortalamadan mutlak hatalarının karekökü ve ortalamadan mutlak hata değerleri bulunmuştur. [6]

Çizelge 5.1: Tahminleme yapan modelin RMSE, Belirtme katsayısı ve MAE değerleri

RMSE	$\boldsymbol{R^2}$	MAE
1,14	0,80	0,87

Model yeni gelen bir aracın ortalama yakıt tüketimi değerinin %80'ini açıklayabilecek güce sahiptir.

Öte yandan artıkların RMSE değerinin 0'a yakın olması da tahmin edilen değerlerin yayılımının küçük ve tutarlı olduğunun bir başka göstergesidir. Aynı şekilde MAE değerinin de 0'a yakın olmasından tahmin değerlerinin gerçek değerlerine yakın olduğu anlaşılır.

SONUÇ

Bu çalışma, kullanılan veri setindeki araçların kümelenmesi, veri setindeki araçları yakıt tipine göre sınıflayan bir model kurulması ve veri setindeki araçların ortalama yakıt tüketimini tahminleyen bir model kurulması için yapılmıştır.

Kümeleme yapıldığında araçlar kasa tipi, motor hacmi, yakıt tipi, ortalama yakıt tüketimi, karbondioksit değerlendirme puanı ve şanzıman tipine göre 4 kümeye ayrılmış ve kümeler yorumlanmıştır.

Yakıt tipine göre araçları sınıflaması için karar ağacı modeli kurulmuştur. Kurulan bu karar ağacı modelinin doğruluk değeri %86'dır.

Ortalama yakıt tüketiminin tahminlenmesi için karar ağacı modeli kurulmuştur. Kurulan bu modelin belirtme katsayısı %80'dir.

KAYNAKÇA

Ayşegül Ö. (2009). SVM Classification for Imbalanced Datasets with Multi Objective Optimization Framework (Yüksek Lisans Tezi). https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=NX16tLMoCJ3Jy6Y08pQV2Q&no=exSvbopvPfshu5BktHxO9g

Bozdoğan, H. (2004) *Statistical Data Mining and Knowledge Discovery*, Chapman & Hall.

Erar, A.& Gülay Kıroğlu (2007)*Veri Çözümleme Ders Notları*, Nobel Akademi Yayınları.

Kıroğlu (Başarır), Gülay (2001) *Uygulamalı Parametrik Olmayan İstatistiksel Yöntemler*, Paymaş, İstanbul.

Nazmiye Yalçın, Kümeleme Analizi ve Uygulaması, (Yüksek Lisans Tezi) https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=qxA-Gbq_PI3wRLbvriZomQ&no=AFRDdt1cP89_g3vRrfOO1w

[1]https://dergipark.org.tr/tr/download/article-file/588956#:~:text=Ward%20y%C3%B6ntemi%2C%20aglomeratif%20k%C3%BCmeleme%20y%C3%B6ntemleri,gh%20ve%20Legendre%2C%202014).

[2] https://en.wikipedia.org/wiki/K-means_clustering

[3]https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

[4]https://www.educative.io/answers/what-is-decision-tree-pruning-and-how-is-it-done

[5]https://bookdown.org/gmli64/do_a_data_science_project_in_10_days/prediction-with-decision-trees.html

[6]https://www.veribilimiokulu.com/r-kare-ve-duzeltilmis-r-kare/