

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SENSÖR IOT VERİLERİ ÜZERİNDE KEŞİFSEL VERİ
ANALİZİ ve DEĞİŞKEN SEÇİMİ UYGULAMASI

MERVE DUMAN ALMANASSRA
YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

GEBZE
2022

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SENSÖR IOT VERİLERİ ÜZERİNDE
KEŞİFSEL VERİ ANALİZİ ve DEĞİŞKEN
SEÇİMİ UYGULAMASI

MERVE DUMAN ALMANASSRA
YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMANI
PROF. DR. AYHAN DEMİRİZ

GEBZE
2022

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**EXPLANATORY DATA ANALYSIS AND
APPLICATION of VARIABLE ELECTION
ON SENSOR IOT DATA**

MERVE DUMAN ALMANASSRA
A THESIS SUBMITTED FOR THE DEGREE OF
MASTER OF SCIENCE
DEPARTMENT OF INDUSTRIAL ENGINEERING

THESIS SUPERVISOR
PROF. DR. AYHAN DEMİRİZ

GEBZE
2022



YÜKSEK LİSANS JÜRİ ONAY FORMU

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 21/07/2022 tarih ve 2022/35 sayılı kararıyla oluşturulan jüri tarafından 03/08/2022 tarihinde tez savunma sınavı yapılan Merve DUMAN'ın tez çalışması Endüstri Mühendisliği Anabilim Dalında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

JÜRİ

ÜYE

(TEZ DANIŞMANI)

: Prof. Dr. Ayhan Demiriz

ÜYE

: Dr. Öğr. Üyesi Kemal Dinçer DİNGEÇ

ÜYE

: Dr. Öğr. Üyesi Elif Elçin GÜNAY

ONAY

Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun
...../...../..... tarih ve/..... sayılı kararı.

İMZA/MÜHÜR

ÖZET

Sensör IoT verilerinin analizi günümüzde birçok sorunun tespit edilmesi, problemlerin çözümlenmesi ve bilinmeyen hakkında yorum yapmamıza olanak sağlamaktadır. Bu çalışmada cam üretim fırınlarında kullanılan sensörlerdeki veriler ele alınmıştır. Öncelikle sensörlerin konumları tespit edilmiştir. Aynı proseste kullanılan sensörler ele alınarak proses içerisindeki sensörlerin birbiri ile arasındaki ilişki ve etkileşimler incelenmiştir. Bu veriler oldukça karmaşık ve büyük değer içeren verilerdir. Bulut üzerinde tutulan veriler Dataiku programı kullanılarak analiz edilmiştir. Keşifsel Veri Analizi ile verilerdeki anomaliler tespit edilmiş olup istatistiksel ve görsel grafikler vasıtasıyla veriler yorumlanabilir ve analiz edilebilir şekle getirilmiştir. Ayrıca birden fazla sensör arasında ilişki olup olmadığı da korelasyon analizi ile incelenmiştir. Verilerin analizi aşamasından elimizde birçok sensör bulunmaktadır. Tüm sensörlerin analiz edilmesi zaman ve maliyet açısından olumsuz etkiye neden olabilmektedir. Bu durumda sadece bizim için önem arz eden ve daha anlamlı olan sensörlerin tespit edilip analizleri bu sensörler üzerinde yoğunlaştırmak işimizi büyük ölçüde kolaylaştıracaktır. Bu aşamada Dataiku programı üzerinde uygulanan entegrasyon ile R dili ile geliştirilen algoritma kullanılarak kullandığımız değişkenlerden hangilerinin bizim için daha fazla anlamlı olduğu hakkında bir sonuca ulaşılmıştır.

Anahtar Kelimeler: IoT, Keşifsel Veri Analizi, Değişken Seçimi, Zaman Serileri, Korelasyon Analizi.

SUMMARY

Analysis of sensor IoT data allows us to identify problems, fix problems, and comment on unknown phenomena. In this study, the data on the sensors used in glass production furnaces are discussed. First, the locations of the sensors were determined. By considering the sensors used in the same process, the relationships and interactions between the sensors in this process are examined. These data are very complex and contain large volume data. The data stored in the cloud was analyzed using the Dataiku program. With Exploratory Data Analysis, anomalies in the data have been detected and the data have made interpretable and analyzable using statistical analysis and visual graphics. Additionally, whether a relationship between more than one sensor was examined by correlation analysis. We have many sensors at our disposal from the analysis stage of the data. Analyzing all sensors can have adverse effects in terms of time and cost. Here, detecting only the important sensors is more meaningful to us and concentrating the analysis on these sensors will greatly facilitate our work. At this arrange, a conclusion has come to approximately which factors we utilized were more significant for us by using the integration connected to the Dataiku program and the calculation created with the R language.

Key Words: IoT, Exploratory Data Analysis, Variable Selection, Time Series, Correlation Analysis.

TEŞEKKÜR

Yüksek lisans eğitimim ve tez çalışmalarım boyunca devamlı beni destekleyen ve her zaman yanımda olan danışmanım Prof. Dr. Ayhan DEMİRİZ'e

Ve tez çalışmalarım boyunca tecrübelerinden yararlandığım çalışma arkadaşlarım Seçil KÖKTÜRK ve Fatma BAKIŞ'a teşekkür ve minnetlerimi sunarım.

İÇİNDEKİLER

	<u>Sayfa</u>
ÖZET	v
SUMMARY	vi
TEŞEKKÜR	vii
İÇİNDEKİLER	viii
SİMGELER ve KISALTMALAR DİZİNİ	xi
ŞEKİLLER DİZİNİ	xii
TABLolar DİZİNİ	xiv
1. GİRİŞ	1
2. KEŞİFSEL VERİ ANALİZİ	2
2.1. Keşifsel Veri Analizi'nin Felsefesi	2
2.2. Keşifsel Veri Analizi'nin Amacı	3
2.3. Keşifsel Veri Analizi'nin Adımları	4
2.3.1. Gerekli Verilerin Toplanması	5
2.3.1.1. Veri Matrisi ve Değişken Tipleri	5
2.3.1.2. Eksik Değerler	5
2.3.1.3. Aykırı Değerler	6
2.3.1.4. Tipik Değerler	7
2.3.2. Veri Setinin Analiz Edilmesi	7
2.3.2.1. Özet Tablolar	7
2.3.2.2. Merkezi Eğilim	8
2.3.2.3. Varyasyon	10
2.3.2.4. Kovaryasyon	10
2.3.2.5. Desen ve Model	11
2.3.2.6. Veri Setinin Dağılımının Gözlenmesi	11
2.3.3. Veri Görselleştirme	11
2.3.3.1. Otokorelasyon Grafiği	12
2.3.3.2. Kutu Grafiği	13
2.3.3.3. Histogram	16
2.3.3.4. Dal ve Yaprak Grafiği	17

2.3.3.5. Gecikme Grafiđi	18
2.3.3.6. Dađılım Grafiđi	19
2.3.3.7. Q-Q Grafiđi	20
2.3.3.8. Isı Haritaları ve Yüzey Grafikleri	21
2.3.4. Korelasyon ve Kovaryans	22
2.3.5. Boyut Analizi	24
2.3.5.1. Temel Bileşen Analizi	24
2.3.6. Kümeleme	26
2.3.7. Deđişken Seçimi	27
2.3.7.1. Filtreleme Yöntemleri	28
2.3.7.2. Sarmalayıcı Yöntemler	30
2.3.7.3. Gömülü Yöntemler	31
2.3.7.4. Grafiksel Modeller	32
2.3.7.5. DiscoVars Yaklaşımı	33
3. SAYISAL SONUÇLAR ve YORUMLAR	36
3.1. Analizde Kullanılan Programlar	37
3.1.1. Google Cloud Platform (GCP)	37
3.1.2. Dataiku	37
3.1.3. RStudio	37
3.2. Veri Kümesinin Detayları	37
3.3. Veri Kümesi Dönüşümleri ve Standardizasyon	38
3.3.1. Veri Kaynađından Verilerin Çekilmesi	38
3.3.2. Veri Frekansının Düzenlenmesi	39
3.3.3. Veri Formatının Düzenlenmesi	39
3.3.4. Eksik Verilerin Hariç Tutulması	39
3.4. Veri Özet Tablosunun Oluşturulması	39
3.5. Aykırı Deđerlerin Tespit Edilmesi	40
3.6. Otokorelasyonun Hesaplanması	41
3.7. Veri Dađılımlarının İncelenmesi	42
3.8. Korelasyon Matrisi	43
3.9. Deđişken Seçimi ve Kümeleme Algoritmasının Kullanılması	44
4. SONUÇ	49

KAYNAKLAR	50
ÖZGEÇMİŞ	53
EKLER	54

SİMGELER VE KISALTMALAR DİZİNİ

Simgeler ve Açıklamalar

Kısaltmalar

PCA	:	Temel Bileşen Analizi(Principal Component Analysis)
GTÜ	:	Gebze Teknik Üniversitesi
lm	:	Lineer Model
GCP	:	Google Cloud Platform
IoT	:	Nesnelerin İnterneti (Internet of Things)
EDA	:	Keşifsel Veri Analizi,(Exploratory Data Analysis)

ŞEKİLLER DİZİNİ

<u>Sekil No:</u>	<u>Sayfa</u>
2.1: Tipik veri bilimi iş akışı.	4
2.2: Dağılım türleri.	9
2.3: EDA teknikleri.	12
2.4: Otokorelasyon grafiği.	12
2.5: Kutu grafiği.	14
2.6: Kutu grafiği aykırı değerler örneği.	15
2.7: Histogram grafiği.	16
2.8: Dal ve yaprak grafiği.	18
2.9: Gecikme grafiği.	18
2.10: Dağılım grafiği.	19
2.11: Q-Q grafiği.	20
2.12: Isı haritaları ve yüzey grafikleri.	21
2.13: Kovaryans örnekleri.	23
2.14: Korelasyon matrisi.	24
2.15: PCA grafik gösterimi.	25
2.16: Kümeleme grafiği.	27
2.17: Markov zinciri.	32
2.18: Bayes teoremi.	33
3.1: Cam soğutma gösterimi.	36
3.2: Veriyi elde etme ve analiz yol haritası.	36
3.3: Örnek kod ekran görüntüsü.	40
3.4: Kutu grafiği aykırı verileri hesaplama kodu.	40
3.5: SENSOR600 kutu grafiği.	41
3.6: SENSOR600 otokorelasyon grafiği.	42
3.7: SENSOR600 histogram grafiği.	42
3.8: Korelasyon matrisi.	43
3.9: Sensörlerin tahmini konumları.	44
3.10: Lasso ve Eigen value kullanılarak geliştirilen grafiksel model.	45
3.11: K-means kümeleme algoritması.	46
3.12: Lasso ve Alpha ölçütü kullanılarak geliştirilen grafiksel model.	47

TABLolar DİZİNİ

<u>Tablo No:</u>	<u>Sayfa</u>
2.1: Geleneksel EDA ve modern EDA karşılaştırması.	3
2.2: DiscoVars algoritması.	34
3.1: Sensör verileri detayları.	38
3.2: Veri özet tablosu.	40
3.3: Sensör verilerindeki aykırı değerler.	41
3.4: Lasso ve Eigen ölçütü kullanılarak bulunan değişkenler.	45
3.5: Lasso ve Alpha ölçütü kullanılarak bulunan değişkenler.	47

1. GİRİŞ

Endüstri 4.0 ile birlikte endüstrilerde IoT sensör kullanımları artmıştır. Bu sensörler tarafından üretilen veriler gün geçtikçe hacim kazanmaktadır. IBM'e göre her gün 2,3 milyar GB veri üretiliyor ve 2 yılda bir veri üretim oranı ikiye katlanıyor. Endüstriler bu verilerden fayda sağlamak amacıyla verileri topluyor, temizliyor ve analiz ediyor ama bu süreçte oluşan maliyetlerden de kaçınmıyor. PWC'ye göre iş yatırımları, 2020'deki genel IoT harcamalarının %50'sinden fazlasını oluşturdu [1]. IDC'nin yaptığı bir çalışmada IoT üretim sektöründe fazlaca kullanılıyor ve tüm sektör harcamalarının yarısından fazlasını oluşturuyor [2]. Rekabete ayak uydurmak için elde edilen verilerden mevcut sorunlara çözüm bulmanın yanı sıra gizli kalmış değerleri keşfetmek bu noktada önem arz ediyor. Keşifsel Veri Analizi, verilerimizde gizlenmiş olan bilgilere ulaşmak için oldukça yaygın kullanılan yöntemlerden birisidir. Bu yöntem şirketlerin faaliyetlerini optimize etmek, verimliliklerini arttırmak ve karar verme sürecinde destekleyici olmak için kullanılır.

Keşifsel Veri Analizi, görselleştirme araçlarını kullanarak kullanıcıların anlayışını geliştirmeyi ve yeni bilgileri etkili bir şekilde iletmeyi hedefler. Ayrıca veriler arasındaki ilişkileri ortaya çıkarır. Bu süreçte sensörlerdeki verilerde analizlerimizi saptıracak ve yanlış kararlar almamızda rol oynayacak etkenleri tespit edilmesi de büyük önem arz etmektedir. Veriler eksik, aykırı veya değişken tipleri yanlış olabilir. Bunların elimine edilmesi doğru bir yaklaşımdır. Veri analizi yaparken elimizdeki değişkenlerin miktarı bizi zaman-maliyet açısından etkileyebilmektedir. Değişkenler arasından hangilerinin bizim için önemli olduğunu belirleyerek analize devam etmek, az öneme sahip değişkenler üzerinde kaybettiğimiz zamanı ve parayı bize geri kazandıracaktır.

Bu çalışmanın amacı verilerdeki gizli alması değerleri keşfetmek, verilerimizin kalitesini artıracak sorular ve desenleri elde etmek ve önemli değişkenleri belirleyerek maliyet azaltmaktır.

2. KEŞİFSEL VERİ ANALİZİ

Veri, yapılan gözlemler sonucunda bilgi sağladığımız hammadDEDİR. Veri analizi ise bu verilerden öngörü ve yorumlar geliştirmemizi sağlar. Bu nedenle problemlerin çözümünde ve faydalı bilgileri keşfetmemizde keşifsel veri analizi büyük bir önem taşımaktadır. Keşifsel Veri Analizi (Exploratory Data Analysis,EDA), görselleştirme araçları yardımıyla çeşitli grafikler oluşturarak verileri analiz etme yöntemidir. Bu yöntem istatistiksel çalışmalarda oldukça fazla kullanılmaktadır. İlgili çalışmamızda bir modelleme yapacak olsak da olmasak da bu analizi yapmamız veriyi anlamak açısından önemlidir. Keşifsel Veri Analizi'nin tanımlayıcısı olan Tukey fizik ve sosyal bilimlere önemli katkılarda bulunmuştur. Ayrıca Tukey, veri analizi çalışmalarına büyük ölçüde katkı sağlamış ve bizlere ışık tutmuştur. Keşifsel Veri Analizi vasıtasıyla betimsel istatistikleri yeniden canlandırmıştır ve paradigmasını değiştirmiştir. Tukey veri analizini, “Verileri analiz etme prosedürleri, bu tür prosedürlerin sonuçlarını yorumlama teknikleri, analizini daha kolay, daha kesin veya daha doğru hale getirmek için veri toplamayı planlama yolları ve verilerin analizi için geçerli olan (matematiksel) istatistiklerin tüm makineleri ve sonuçları.” olarak tanımlamıştır [3].

Keşifsel Veri Analizi açık uçlu ve yinelemeli analiz yaklaşımıdır. Önceden tasarlanmış olan kavram ve kurallardan bağımsız olarak verileri incelememize olanak sağlar. Bunun neticesinde bize iyileştirme önerileri sunar. Behrens'e göre Keşifsel Veri Analizi, Kullanıcının beklenmedik örüntüleri gözlemlemesine ve bu örüntülerin etkisini düşünmesine izin veren bir süreçtir [4].

2.1. Keşifsel Veri Analizi'nin Felsefesi

Keşifsel Veri Analizi'nin temel yaklaşımı yaratıcı bir süreç olmasıdır. Yani elimizdeki veriyi keşfetmemiz adına bize sorular sorar. Bu sorular içerisinde en önemli soruyu bulmamıza olanak sağlar. Bunu yapmasındaki amaç analizimize başlarken hangi soruları soracağımızı bilmediğimizdendir. Her defasında doğru soruları yönelterek analizimize farklı bir bakış açısı kazandırırız ve faydaya bir adım daha yaklaşmış oluruz. Buradaki amacımız ise analizimizi yaparken nasıl bir model

kullanmalıyız veya nasıl bir grafik ile verilerimizi daha iyi açıklarız gibi sorulara cevap aramaktır. Bu süreci yönetmenin doğru veya yanlış bir yolu yoktur.

2.2. Keşifsel Veri Analizi'nin Amacı

Keşifsel Veri Analizi'nin amaçları arasında

- Gözlenen olayların nedenleri hakkında önerilerde bulunmak,
- Varsayımlarda bulunmak,
- Uygun istatistiksel araç ve teknik seçimlerini desteklemek,
- Gözlenen olayın hakkında önerilerde bulunmak,
- Varsayımlarda bulunmak ve bunları değerlendirmek,
- Daha fazla veri elde etmek için bilgi sağlamak [4],
- Aykırı değerleri ve anomalileri tespit etmek,
- Alttaki yapıyı ortaya çıkartmak,
- Önemli değişkenleri ayıklamak [5],

gibi maddeler bulunmaktadır. Keşifsel Veri Analizinde elimizde belirli bir kural seti bulunmaz. Elinizdeki veriye göre çıktıları şekillendirirsiniz. Buradaki asıl amaç kaliteyi yükseltmektir. Geleneksel araştırma yöntemlerini takip etmek yerine, veri bilimi, anlık soruları yanıtlamak için sürekli ve büyük miktarda verinin analizine gerçekleştirir.

Tablo 2.1: Geleneksel EDA ve Modern EDA karşılaştırması.

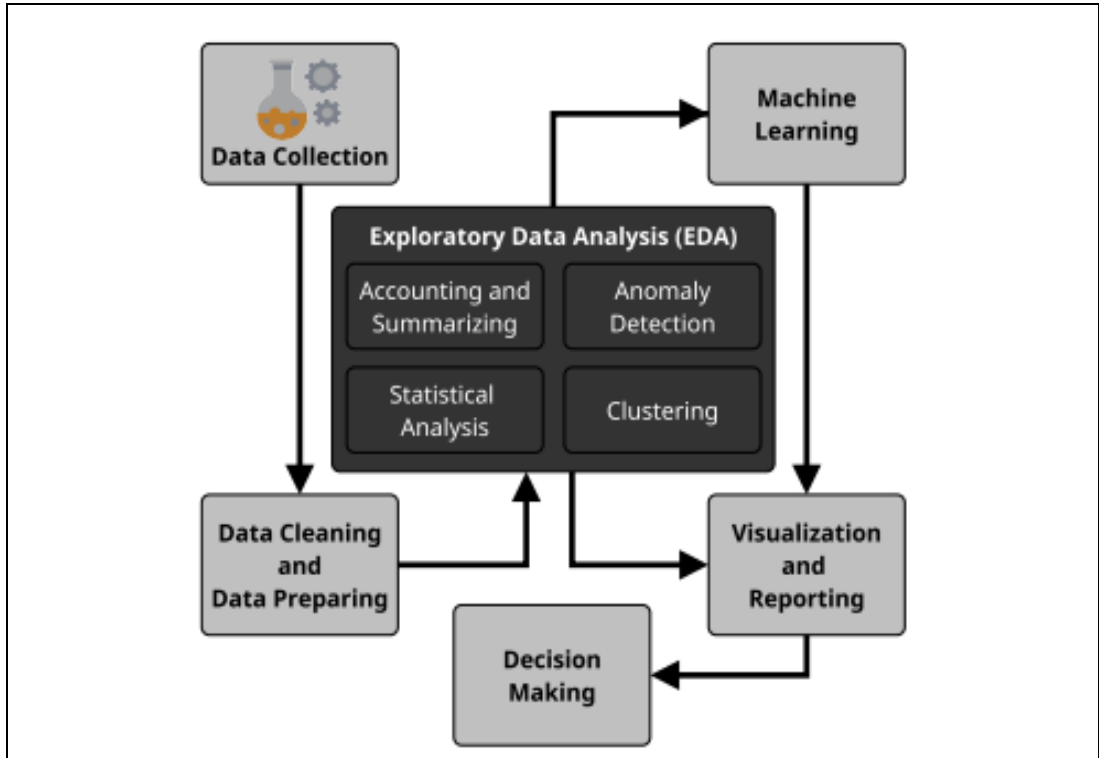
	Geleneksel EDA	Modern EDA
Kapsam	Görselleştirilmiş keşifsel analiz	Genel keşifsel analiz
Araçlar	Basit aritmetik işlemler Çizimi kolay resimler	Betimsel analiz Gelişmiş görselleştirme teknikleri Veri madenciliği teknikleri
Ana Özellikler	Kesin cevaplar yok	Kesin cevaplar sağlayabilir

2.3. Keşifsel Veri Analizi'nin Adımları

Keşifsel veri analizinde tamamlanması gereken altı temel adım aşağıdaki gibidir.

- Veri kümesinin veri özelliklerinin değerlendirilmesi,
- Veri kalitesinin, eksik ve aykırı değerlerin değerlendirilmesi,
- Nitelikler veya nitelik kombinasyonları arasındaki örüntülerin veya ilişkilerin ilk tespiti,
- Tüm gözlemlerin aynı temel popülasyondan gelip gelmediğinin belirlenmesi,
- Analizler sonucunda verilerimize uygun olarak yönelttiğimiz sorulara uygun yanıtları aramak,
- Elde ettiğimiz cevaplar ile sorunun düzeltilmesi veya yeni soruların üretilmesi

olarak sayılabilir [6].



Şekil 2.1: Tipik veri bilimi iş akışı.

2.3.1. Gerekli Verilerin Toplanması

Keşifsel Veri Analizi sürecine başlamanın en önemli adımı gerekli verileri elde etmek ve bu verileri hazırlamaktır. Analizciler çok çeşitli ve yüksek boyutlu veriye sahip olabilirler. Bu verileri toplamak için uzun zaman ayırırlar ve çeşitli araçlar kullanırlar. Toplanan veriler farklı görünebilir. Bu aşamada verilerin temiz ve yorumlanabilir olması çok önemlidir. Bunun için çaba sarf edilmelidir. Gerekli verileri toplayıp düzenini sağladıktan sonra, veri setinde yer alan değişkenler incelenerek sürece başlanabilir [7].

Hadley Wickham (2014) temiz ve düzenli veriler için bir yol haritası önermiştir. Kendisine göre düzenli verileri üç madde ile özetlemek mümkündür.

- Her değişken bir sütun oluşturur.
- Her gözlem bir satır oluşturur.
- Her gözlem birimi bir tablo oluşturur [8].

olarak sayılabilir.

2.3.1.1. Veri Matrisi ve Değişken Tipleri

Tablo verileri n gözlemden ve p ölçümden oluşan bir veri setidir. Veriler $n \times p$ şeklinde bir veri matrisi oluştururlar. Satırlar örnekler, durumlar, nesneler, birimler, kayıtlar veya gözlemler ($i = 1, \dots, n$), sütunlar ise değişkenler veya alanlar ($j = 1, \dots, p$) olarak ifade edilebilir [9].

2.3.1.2. Eksik Değerler

Eksik değerleri olan bir değişkeni dikkatli olarak incelemek gereklidir. Eğer tüm sütunda eksik değer mevcut ise bu sütun bizim için bir anlam ifade etmeyecektir. Eksik değerler tahminler ve analizler üzerinde oldukça olumsuz etkiye sahiptirler. Bu tür değişkenler analizden çıkarılabilir. Bazı durumlarda ise bu eksik değerlerin deseni bize veri seti hakkında önemli bilgiler veriyor olabilir.

2.3.1.3. Aykırı Değerler

Aykırı değerler “Bir veri setinin geri kalanıyla tutarsız görünen bir gözlem veya bu gözlemlerin alt kümesi” olarak tanımlanmıştır. (Barnett ve Lewis 1997) Burada önemli olan nokta bu gözlemin bir popülasyona ait olup olmadığıdır. Değişkenin aykırı değer olup olmadığına yönelik olasılık yaklaşımı normal dağılımın özelliklerine dayanmaktadır [9].

Aykırı değerler olağan dışı gözlemlerdir. Veri deseninize uymayan veri noktalarıdır. Bu noktalar bazen veri girişi sırasında bazen de veri aktarımı sırasında ortaya çıkabilir. Çok fazla veriye sahip olduğumuzda bunları tespit etmek zor olabilir. Analizlerimizi aykırı değerler varken ve yokken tekrarlamak iyi bir uygulamadır. Eğer bu değerler verilerimiz üzerinde minimum etkilere sahiplerse ve neden orada oldukları hakkında bilgimiz yoksa bu verileri eksik veriler ile değiştirebiliriz. Eğer bu veriler veri setimiz üzerinde oldukça fazla önemli bir etkiye sahiplerse bu durumda buna neyin sebep olduğunu anlamamız ve setimizden çıkartmamız en doğrusu olur [10].

$$x_{ij} = \begin{cases} \text{aykırı değer,} & \text{eğer } |x_{ij} - c_j|/s_j > t \\ \text{aykırı olmayan değer,} & \text{diğer} \end{cases} \quad (2.1)$$

x_{ij} – örnek ortalama veya medyan

s_j – örnek standart sapma veya ortalamadan mutlak sapma

t – pozitif sayı

Aykırı değerler ile karşılaştığımızda aşağıdaki iki seçeneği değerlendirmeliyiz.

- Aykırı verilerin bulunduğu tüm satırı çıkartmak. Bu seçenek çok önerilmez çünkü bir ölçümün geçersiz olması diğer ölçümlerinde geçersiz olduğu anlamına gelmez. Eğer veri kaliteniz düşük ise ve bu yaklaşımı tüm değişkenlere uygularsanız elinizde hiç veri kalmayabilir.
- Aykırı verileri eksik veriler ile yer değiştirmek.

Aykırı değerleri sınıflandırmak için winsorize denilen bir yöntemi kullanabiliriz. Bu yöntem veri setini ortalamaya göre ayarlar [9].

2.3.1.4. Tipik Değerler

Grafiklerde bulunan sütunlar değişkenlerde en çok kullanılan değerleri veya daha az yaygın olarak kullanılan değerleri göstermektedir. Eğer bir yerde sütun göremiyorsak bunun anlamı verilerimizde görülmeyen değerlerin bulunmasıdır. Bu gibi durumları Keşifsel Veri Analizi'nin genel felsefesinde de anlattığımız gibi faydaya dönüştürmeli ve bu verilerde beklenmedik bir şey olup olmadığına bakmalıyız.

Bu gibi durumlarda aşağıdaki soruları sormamız faydalı olur.

- Hangi veriler en yaygındır?
- Bu veriler beklentilerimizi karşılıyor mu?
- Olağandışı desenler görebiliyor muyuz? Bunları nasıl açıklayabiliriz?
- Eğer benzer değer kümelerimiz var ise bu durum verilerimizde alt grupların bulunduğu anlamına gelmektedir. Bu durumu anlamak için şu soruları sorabiliriz.
- Her küme için gözlemler birbirine nasıl benziyor?
- Ayrı kümelerin gözlemlerinin birbirinden farkı nasıldır?
- Kümeleri nasıl açıklayabilir veya tanımlayabiliriz.
- Kümelerin görünümü niçin yanıltıcı olabilir?

Bu ve bunun gibi sorular değişkenler arasındaki ilişkileri keşfetmemiz ve gözlemlememizde yardımcı olacaklardır [10].

2.3.2. Veri Setinin Analiz Edilmesi

2.3.2.1. Özet Tablolar

Özet tablolar için tüm değişkenlerimizi listeler ve böylece betimsel istatistikleri hesaplarız. Betimsel istatistikler, puan dağılımının bir özetini sağlar. Genel olarak, bir Keşifsel Veri Analizi süreci ortalama, medyan, mod, aralık ve standart sapmanın hesaplanmasını içermelidir; ancak diğer tanımlayıcı istatistikler daha ileri düzey uygulayıcılar için faydalı olabilir (Martin & Bridgmon, 2012).

Bunlar;

- Ortalama,
- Standart sapma,
- Medyan,
- Minimum ve maksimum,
- Çarpıklık ve basıklık

olarak sayılabilir.

2.3.2.2. Merkezi Eğilim

Bir dağılımın merkezi eğilimi veya “konumu” tipik veya orta değerlerle ilgilidir. Merkezi eğilimin yaygın, yararlı ölçüleri, (aritmetik) ortalama, medyan ve bazen mod olarak adlandırılan istatistiklerdir. Bazen, geometrik, harmonik, kesik veya Winsorized araçlar gibi diğer araçlar, merkezilik ölçüleri olarak kullanılır. Çoğu yazar "ortalama" terimini aritmetik ortalama ile eşanlamlı olarak kullanırken, bazıları ortalamayı geometrik, harmonik ve diğer araçları da içerecek şekilde daha geniş bir anlamda kullanır. x_1 'den x_n 'ye kadar etiketlenmiş n veri değerimiz olduğunu varsayarsak , örnek (aritmetik) ortalamayı hesaplama formülü şekil 2.2 deki gibidir.

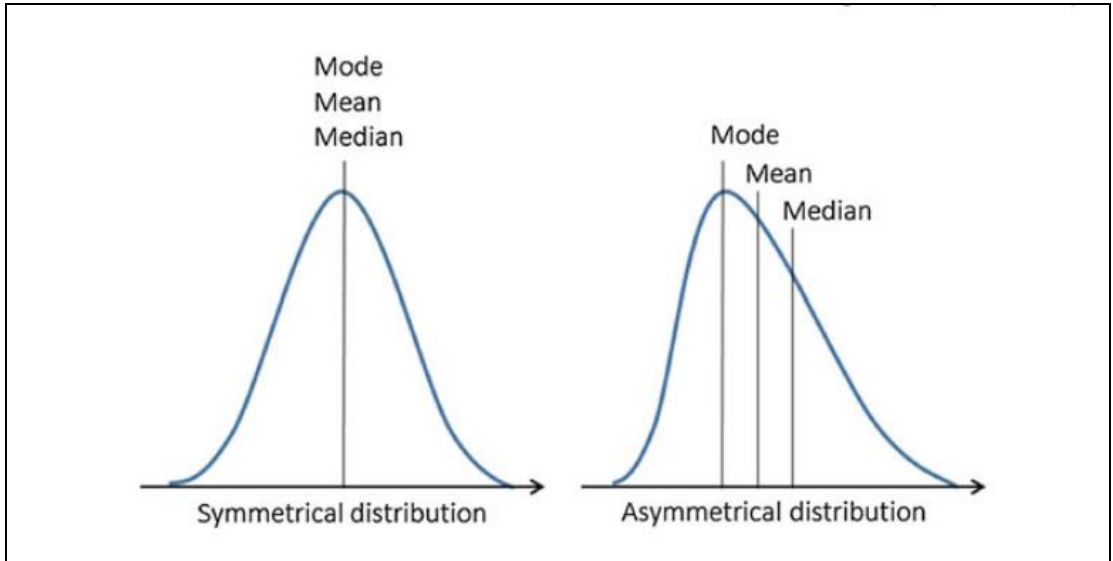
$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (2.2)$$

Aritmetik ortalama, basitçe, tüm veri değerlerinin toplamının, değer sayısına bölümüdür. Simetrik olarak şekillendirilmiş herhangi bir dağılım için ortalama, simetrinin etrafında tutulduğu noktadır.

Medyan, bir başka merkezi eğilim ölçüsüdür. Örnek medyan tüm değerler sıralı bir listeye konduktan sonra ortadaki değer. Çift sayıda değer varsa, ortadaki iki değer ortalamasını alın. Ortada bağ varsa kullanacağımız istatistiksel yazılım tarafından bazı özel ayarlamalar yapılır. Kesikli rasgele değişkenler için olağandışı durumlarda, benzersiz bir medyan olmayabilir. Medyan oldukça sağlamdır, çünkü veri değerlerinin neredeyse tamamının üst yarısını ve/veya alt yarısını medyanı

değiştirmeden medyandan herhangi bir mesafe uzağa taşıyabilirsiniz. Daha pratik olarak, birkaç çok yüksek değer veya çok düşük değer genellikle medyan üzerinde hiçbir etkiye sahip değildir. Nadiren kullanılan bir merkezi eğilim ölçüsü, en olası veya sıklıkla meydana gelen değer olan moddur. Daha yaygın olarak, bir dağılımın tek bir tepe noktası (tek modlu) veya iki veya daha fazla tepe noktası (iki modlu veya çok modlu) olup olmadığını tanımlarken basitçe "mod" terimini kullanırız. Simetrik, tek modlu dağılımlarda mod hem ortalamaya hem de medyana eşittir. Tek modlu, çarpık dağılımlarda mod, medyanın ortalamadan diğer tarafındadır. Çok modlu dağılımlarda ya benzersiz bir en yüksek mod yoktur ya da en yüksek mod, merkezi eğilimi temsil etmeyebilir [11].

Çarpıklık ve basıklık, İki ek yararlı tek değişkenli tanımlayıcı, bir dağılımın çarpıklığı ve basıklığıdır. Çarpıklık bir asimetri ölçüsüdür. Dağılımın ortalama etrafındaki simetriye göre sapmasını gösterir. Basıklık, Gauss şekline göre “tepe noktasının” bir ölçüsüdür. Yani tepe noktalarının normalden sapmasıdır [12].



Şekil 2.2: Dağılım türleri.

Varyans, standart sapma ve çeyrekler arası aralık dahil olmak üzere bir dağılımın yayılmasının bir ölçüsü olarak çeşitli istatistiklerde yaygın olarak kullanılır. Yayılma, merkezden ne kadar uzakta veri değerleri bulma ihtimalimizin olduğunun bir göstergesidir. Standart sapma, varyansın basitçe kareköküdür. Bu nedenle, orijinal verilerle aynı birimlere sahiptir, bu da onu daha yorumlanabilir hale getirmeye yardımcı olur. Örnek standart sapması genellikle s sembolü ile gösterilir.

Teorik bir Gauss dağılımı için, ortalama artı veya eksi 1, 2 veya 3 standart sapmanın olasılığın sırasıyla %68,3, %95,4 ve %99,7'sini içerir ve bu, normal dağılımdan elde edilen gerçek veriler için yaklaşık olarak doğru olmalıdır.

Çeyrekler açıklığı (IQR) tanımlamak için önce çeyrekler kavramlarını tanımlamamız gerekir. Bir popülasyonun veya bir örneğin çeyrekleri, dağılımı veya gözlemlenen verileri dörde bölen üç değerdir. Bu nedenle, verilerin dörtte biri, genellikle Q1 olarak yazılan ilk çeyreğin altına düşer; bir yarısı ikinci çeyreğin (Q2) altına düşer ve dörtte üçü üçüncü çeyreğin (Q3) altına düşüyor. Çeyrekler açıklığı güçlü bir yayılma ölçüsüdür [11].

2.3.2.3. Varyasyon

Varyasyon bir değişkenin değerlerinin bir önceki veya sonrakine göre değişme eğilimine verilen addır. Her bir ölçüm ölçümünden ölçüme değişen az miktarda bir hata içerir. Sürekli değişkenlerde her ölçüm farklı sonuçlar verebilir. Kategorik değişkenlerde aynı şekilde tanımlanabilir. Bunu anlamamanın en iyi yolu değişkenlerin dağılımlarını incelemektir [10].

2.3.2.4. Kovaryasyon

Varyasyon bir değişken içindeki davranışı tanımlarken kovaryasyon değişkenler arasındaki davranışı tanımlamaktadır. Kovaryasyon iki veya daha fazla değişkenin ilişkili bir şekilde birlikte değişme eğilimidir. Kovaryasyonu belirlemenin en kolay yolu iki veya daha fazla değişken arasındaki ilişkiyi görselleştirmektir. Bunun nasıl yapılacağı ilgili değişkenin türüne bağlıdır. Varyasyon belirsizlik yaratan bir olgu iken kovaryasyon bu belirsizliği azaltma eğilimindedir. Kovaryasyon nedensel bir ilişkiden kaynaklanıyorsa (özel bir durum), o zaman ikincinin değerini kontrol etmek için bir değişkenin değerini kullanabilirsiniz [10].

2.3.2.5. Desen ve Model

Verilerinizdeki desenler ilişkiler hakkında ipuçları sağlar. Sistematik bir ilişki var ise veriler bir kalıp gibi görünecektir. Modeller, verilerden kalıpları çıkarmak için bir araçtır.

Bu durumda sorulması gereken sorular aşağıdaki gibidir [10].

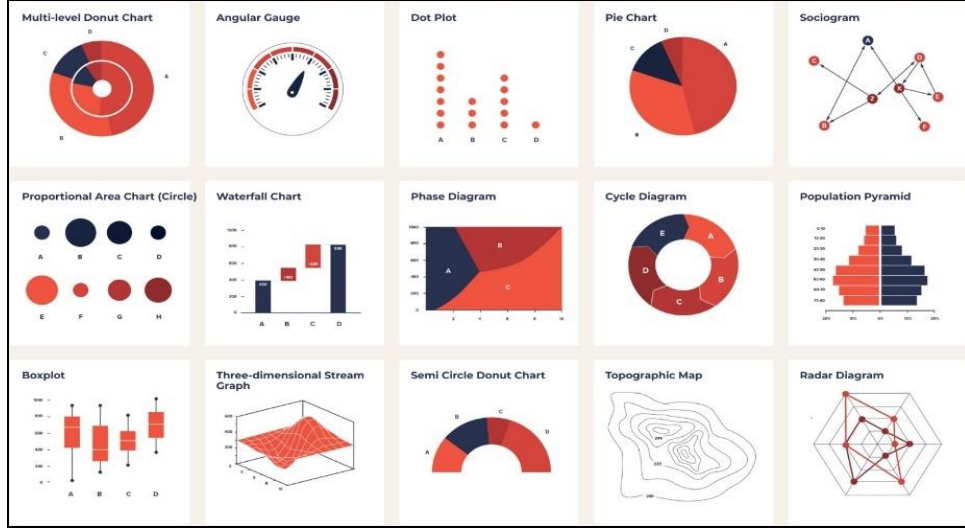
- Bu model tesadüfe bağlı olabilir mi?
- Bu ilişkiyi nasıl tanımlarsınız?
- Bu ilişki ne kadar güçlüdür?
- Başka hangi değişkenler ilişkiyi etkileyebilir?

2.3.2.6. Veri Setini Dağılımının Gözlenmesi

Her bir değişkenin dağılımlarını gözlemlemek için her değişken sütununun istatistiksel özelliklerini incelemek gerekmektedir. Böylece ilgili sapmaları belirlemiş oluruz.

2.3.3. Veri Görselleştirme

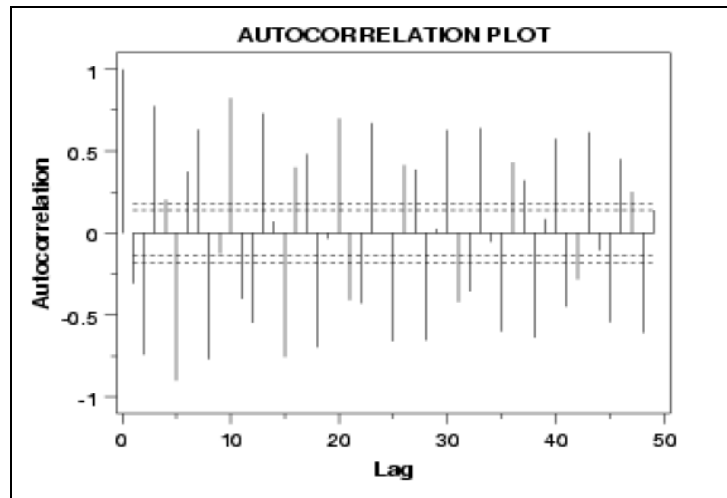
Verilerin görsel olarak araştırılması, anlaşılması için kritik bir öneme sahiptir ve bu görselleştirme Keşifsel Veri Analizi'nin önemli bir parçasıdır. Grafik ekranların, bir kullanıcının veri anlayışını geliştirmeye ve yeni bilgileri etkili bir şekilde iletmesine yardımcı olduğu bilinen bir gerçektir. Bu teknikler, psikoloji ve iş analitiği gibi diğer alanlarda uzun yıllardır yaygın olarak kullanılmaktadır (Alhadad, 2018; Diamond & Mattia, n.d.). Özet istatistikleri sunmak, dağılımları görüntülemek, değişkenler arasındaki ilişkiyi göstermek, model parametrelerinin seçimini kolaylaştırmak, grupları tespit etmek vb. için farklı grafikler kullanılabilir. Ayrıca, grafik biçiminin seçimi, tasvir edilen verinin türüne bağlıdır (Clarke ve diğerleri 2009, 531-534) [13].



Şekil 2.3: EDA teknikleri.

2.3.3.1. Otokorelasyon Grafiği

Otokorelasyon, belirli bir zaman serisi ile bu serinin lag(gecikmeli) hali arasındaki ilişkileri inceler. Otokorelasyon grafikleri, bir veri setindeki rastgeleliği kontrol etmek için yaygın olarak kullanılan bir araçtır. Bu rastgelelik, değişen zaman gecikmelerinde veri değerleri için otokorelasyonların hesaplanmasıyla belirlenir. Rastgele ise, bu tür otokorelasyonlar, tüm zaman gecikmeli ayrımlar için sıfıra yakın olmalıdır. Rastgele değilse, bir veya daha fazla otokorelasyon önemli ölçüde sıfırdan farklı olacaktır.



Şekil 2.4: Otokorelasyon grafiği.

Otokorelasyon grafikleri şu şekilde oluşturulur:

Dikey eksen: Otokorelasyon katsayısını gösterir. R_h -1 ile +1 arasındadır.

$$R_h = C_h / C_0 \quad (2.3)$$

Burada C_h otokovaryans fonksiyonudur.

$$C_h = \frac{1}{N} \sum_{t=1}^{N-h} (Y_t - \bar{Y}) (Y_{t+h} - \bar{Y}) \quad (2.4)$$

Burada C_0 varyans fonksiyonudur.

$$C_0 = \frac{\sum_{t=1}^N (Y_t - \bar{Y})^2}{N} \quad (2.5)$$

Yatay eksen: Gecikme katsayısı(lag) h ($h=1,2,3,\dots$)

Yukarıdaki çizgi ayrıca birkaç yatay referans çizgisi içerir. Orta çizgi sıfırdır. Diğer dört çizgi %95 ve %99 güven aralığıdır.

Otokorelasyon grafiği aşağıdaki sorulara cevap verebilir:

- Veriler rastgele mi?
- Bir gözlem, bitişik bir gözlemle ilgili mi?
- Gözlenen zaman serisi beyaz gürültü mü?
- Gözlenen zaman serisi sinüzoidal midir?
- Gözlenen zaman serisi otoregresif midir?
- Gözlenen zaman serisi için uygun model nedir?
- $Y = \text{sabit} + \text{hata}$ modeli geçerli ve yeterli mi? [5]

2.3.3.2. Kutu Grafiği

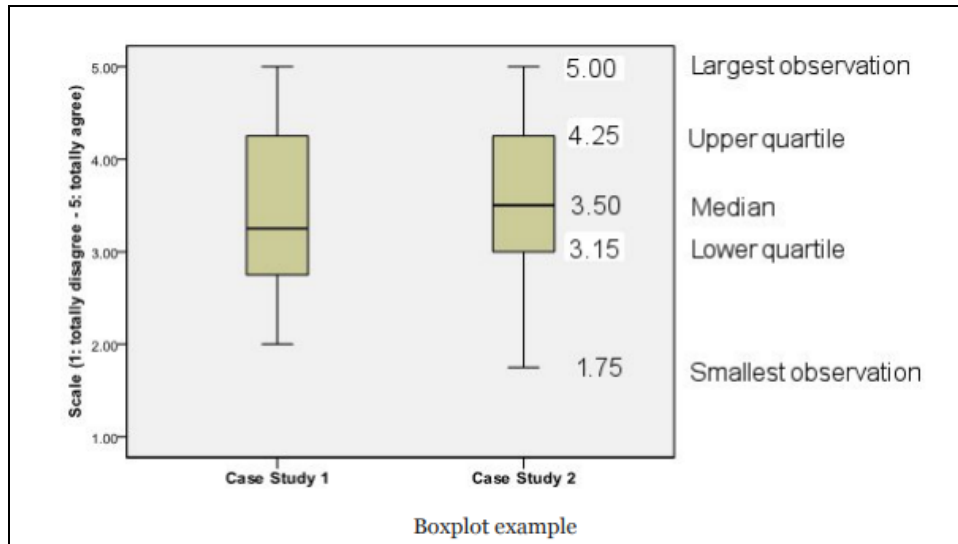
Kutu grafikleri (Chambers 1983), özellikle farklı veri grupları arasındaki konum ve varyasyon değişikliklerini saptamak ve göstermek için, veri kümelerinde konum ve varyasyon bilgilerini iletmek için mükemmel bir araçtır.

En küçük gözlem (örnek minimum), alt çeyrek (%25), ortanca (%50), üst çeyrek (%75) ve en büyük gözlem (örnek maksimum). Aykırı değerler varsa, kutu çizimi de bunları gösterir. Kutu alttan, alt çeyrekten üste, üst çeyreğe kadar inşa edilmiştir.

Bir kutu grafiği elde etmek için;

- Adım 1: Medyanı ve çeyrekleri hesaplayın (alt çeyrek yüzde 25'tir ve üst çeyrek yüzde 75'tir).
- Adım 2: Ortancaya bir sembol çizin (veya bir çizgi) ve alt ve üst çeyrekler arasına bir kutu çizin; bu kutu, verilerin ortadaki %50'sini, yani verilerin "gövdesini" temsil eder.
- Adım 3: Alt çeyrekten minimum noktaya bir çizgi ve üst çeyrekten maksimum noktaya başka bir çizgi çizin. Bu isteğe bağlı olmasına rağmen, tipik olarak bu minimum ve maksimum noktalarda bir sembol çizilir. Böylece kutu grafiği, verilerin ortadaki %50'sini, medyanı ve uç noktaları tanımlar [5].

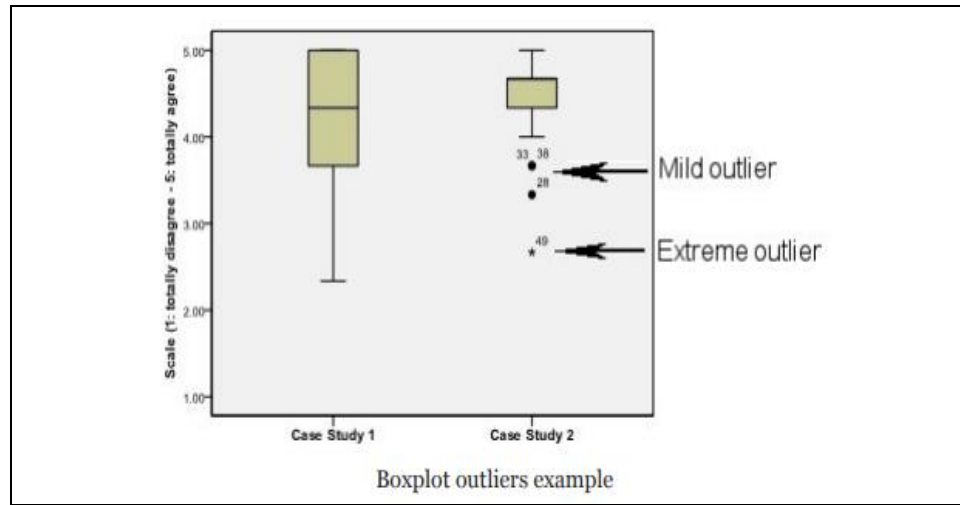
adımları takip edilebilir.



Şekil 2.5: Kutu grafiği.

Kutu grafiğinin, aykırı değerleri daha spesifik olarak tanımlayan yararlı bir varyasyonu vardır. Bu varyasyonu oluşturmak için:

- Medyanı ve alt ve üst değerleri hesaplayın.
- Medyana bir sembol ve alt ve üst çeyrekler arasına bir kutu çizin.
- Çeyrekler arası aralığı (üst ve alt çeyrek arasındaki fark) hesaplayın ve buna IQ diyebiliriz.
- Aşağıdaki noktaları hesaplayın:
 - L1 = alt çeyrek- 1.5*IQ
 - L2 = alt çeyrek- 3.0*IQ U1 = üst çeyrek + 1.5*IQ
 - U2 = üst çeyrek + 3.0*IQ
- Alt çeyrekten minimuma doğru olan çizgi alt çeyrekten L1'den büyük olan en küçük noktaya çizilir. Benzer şekilde, üst çeyrekten maksimuma doğru olan çizgi şimdi U1'den daha küçük olan en büyük noktaya çizilir.
- L1 ile L2 arasındaki veya U1 ile U2 arasındaki noktalar küçük daireler olarak çizilir. L2'den küçük veya U2'den büyük noktalar büyük daireler olarak çizilir.



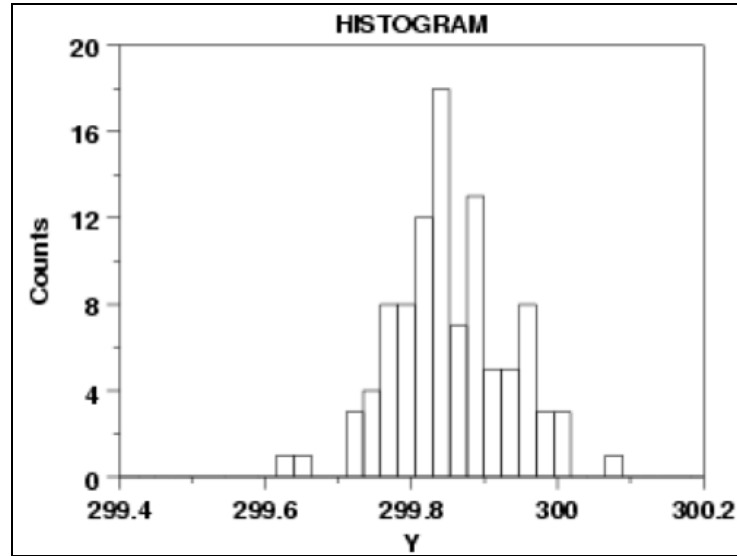
Şekil 2.6: Kutu grafiği aykırı değerler örneği.

Kutu grafiği aşağıdaki sorulara cevap aramak için kullanılabilir.

- Bir faktör önemli midir?
- Konum, alt gruplar arasında farklılık gösteriyor mu?
- Varyasyon alt gruplar arasında farklılık gösteriyor mu?
- Aykırı değerler var mı? [5]

2.3.3.3. Histogram

Histogramın en yaygın biçimi, veri aralığının eşit boyutlu kutulara (sınıflar olarak adlandırılır) bölünmesiyle elde edilir. Daha sonra her bölme için, veri kümesinden her bölmeye düşen noktaların sayısı sayılır. Sınıflar, kullanıcı tarafından keyfi olarak veya bazı sistematik kurallar aracılığıyla tanımlanabilir. Teorik olarak türetilmiş bir dizi kural Scott tarafından önerilmiştir (Scott 1992). Kümülatif histogram, dikey eksenin yalnızca tek bir bölme için sayıları vermekle kalmayıp, bunun yerine o bölme için sayıları ve yanıt değişkeninin daha küçük değerleri için tüm bölmeleri verdiği histogramın bir varyasyonudur.



Şekil 2.7: Histogram grafiği.

Hem histogram hem de kümülatif histogram, sayıların normalleştirilmiş sayımlarla değiştirildiği ek bir değişkene sahiptir. Bu varyantların adları, görelî histogram ve görelî kümülatif histogramdır.

Sayıları normalleştirmenin iki yaygın yolu vardır. Bunlar normalleştirilmiş sayım ve normalleştirilmiş sayı olarak ifade edilebilir. Normalleştirilmiş sayım, bir sınıftaki sayının toplam gözlem sayısına bölümüdür. Bu durumda, bağıl sayımlar bire (veya yüzde ölçeği kullanılıyorsa 100'e) toplanacak şekilde normalleştirilir. Bu, histogram çubuğunun yüksekliğinin her sınıftaki verilerin oranını temsil ettiği sezgisel bir durumdur. Normalleştirilmiş sayı, sınıftaki gözlem sayısının sınıf genişliğinin çarpımına bölünmesiyle elde edilen sayıdır. Bu normalleştirme için

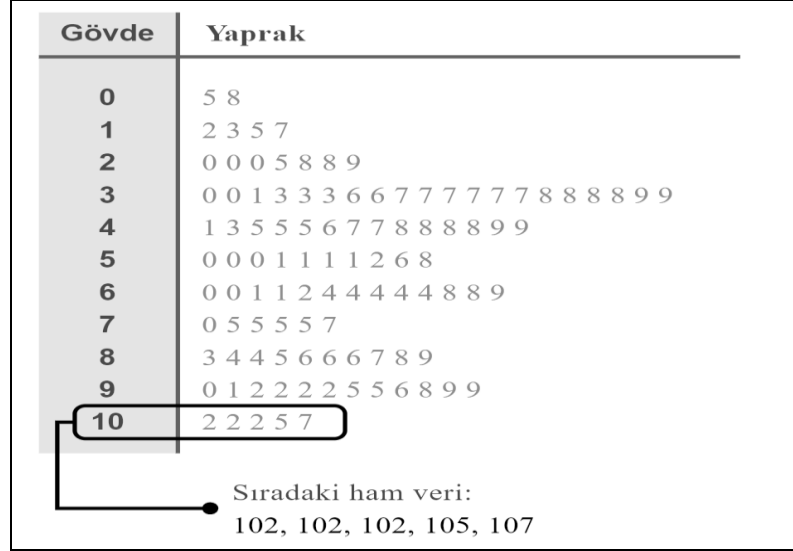
histogramın altındaki alan bire eşittir. Olasılıksal bir bakış açısından, bu normalleştirme görelî bir histogramla sonuçlanır. Bu, olasılık yoğunluk işlevine en çok benzeyen ve kümülatif dağılım işlevine en çok benzeyen görelî kümülatif histogramdır. Histogramın üstüne bir olasılık yoğunluğu veya kümülatif dağılım işlevi yerleştirmek istiyorsanız, bu normalleştirmeyi kullanabilirsiniz. Bu normalleştirme daha az sezgisel olsa da (1'den büyük görelî frekanslara oldukça izin verilir), bir olasılık yoğunluk fonksiyonunu modellemek için histogram kullanıyorsanız, bu uygun normalleştirmedir.

Histogram aşağıdaki soruları cevaplamak için kullanılabilir:

- Veriler ne tür bir nüfus dağılımından geliyor?
- Veriler nerede bulunur?
- Veriler ne kadar yayılmış?
- Veriler simetrik mi yoksa çarpık mı?
- Verilerde aykırı değerler var mı? [8]

2.3.3.4. Dal ve Yaprak Grafiği

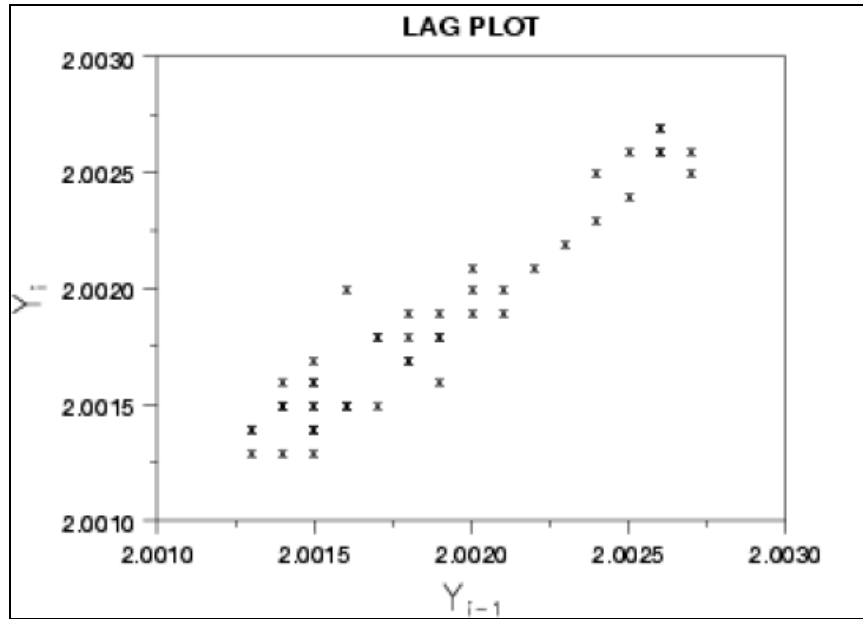
Bir histogramın basit bir ikamesi, bir dal ve yaprak grafiğidir. Bir dal ve yaprak grafiğini elle yapmak bazen bir histogramdan daha kolaydır ve herhangi bir bilgiyi gizlememe eğilimindedir. Bununla birlikte, bir histogramın genellikle bir numune dağılımının şeklini takdir etmek için dal ve yaprak grafiğinden daha iyi olduğu düşünülür. Bir dal ve yaprak grafiği, tüm veri değerlerini ve dağılımın şeklini gösterir [14].



Şekil 2.8: Dal ve yaprak grafiği.

2.3.3.5. Gecikme Grafiği

Bir gecikme grafiği, bir veri kümesinin veya zaman serisinin rastgele olup olmadığını kontrol eder. Rastgele veriler, gecikme grafiğinde tanımlanabilir herhangi bir yapı sergilememelidir. Gecikme grafiğinde rastgele olmayan yapı temel alınan verilerin rastgele olmadığını gösterir [5].



Şekil 2.9: Gecikme grafiği.

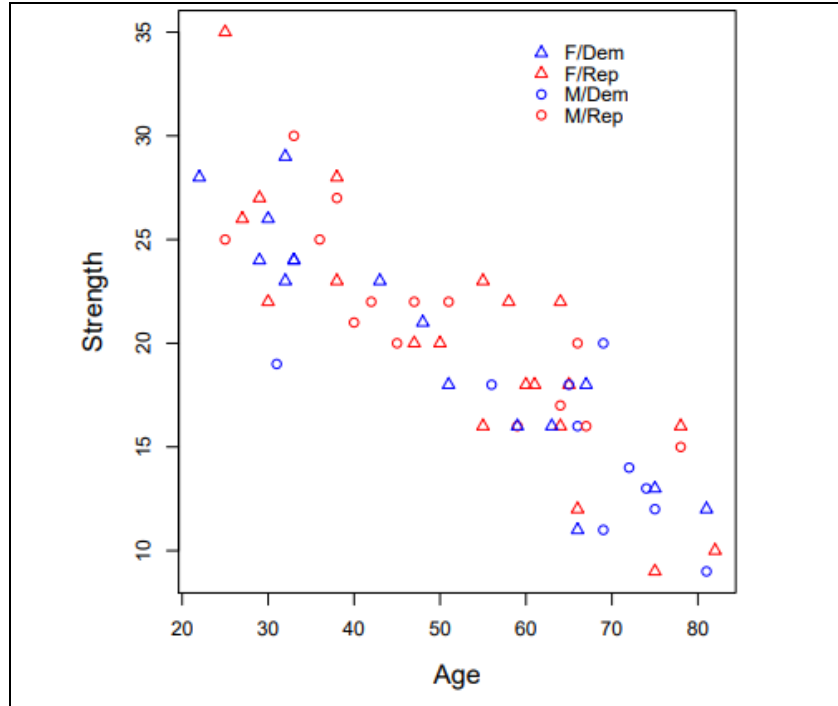
Gecikme, sabit bir zaman yer deęiřtirmesidir. Örneęin, bir veri seti verildięinde $Y_1, Y_2, \dots, Y_n$, Y_2 ve Y_7 , $7 - 2 = 5$ olduęundan gecikme 5'e sahiptir. En yaygın kullanılan gecikme 1 olmasına raęmen, herhangi bir rastgele gecikme için gecikme grafikleri oluřturulabilir. Gecikme grafięi, Y_i 'ye karřı Y_{i-1} deęerlerinin bir grafięidir. Burada tüm i deęerleri için dikey eksen Y_i Yatay eksen ise Y_{i-1} olarak tanımlanır.

Bir gecikme grafięinde ařaęıdaki sorulara yanıt ararız.

- Veriler rastgele mi?
- Verilerde korelasyon var mı?
- Veriler için uygun model nedir?
- Verilerde aykırı deęerler var mı? [8]

2.3.3.6. Daęılım Grafięi

Bir daęılım grafięi, iki deęiřken arasındaki iliřkileri veya iliřkiyi ortaya ıkarır. Bu tür iliřkiler, olay örgüsünde rastgele olmayan herhangi bir yapı ile kendini gösterir (Chambers 1983).



řekil 2.10: Daęılım grafięi.

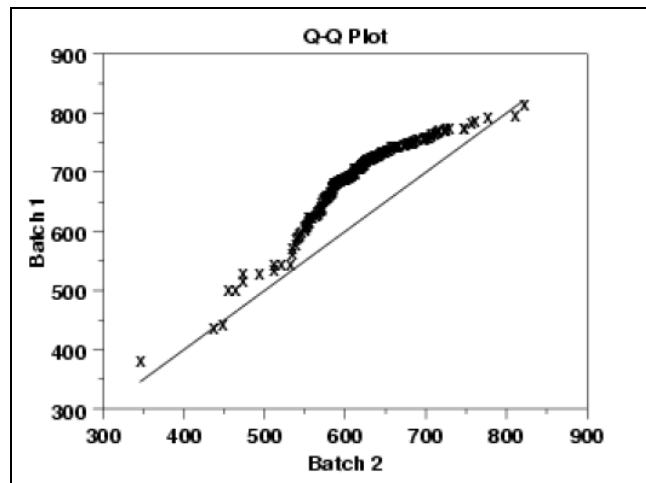
Bir dağılım grafiği, Y değerlerine karşılık X'in karşılık gelen değerlerinin grafiğidir. Dikey eksen Y eksenini olarak tanımlanır ve genelde yanıt değişkenidir. Yatay eksen X eksenini olarak tanımlanır ve genellikle yanıtla ilgili olabileceğinden şüphelendiğimiz bazı değişkenlerdir.

Dağılım grafikleri aşağıdaki sorulara cevap verebilir:

- X ve Y değişkenleri ilişkili mi?
- X ve Y değişkenleri doğrusal olarak ilişkili mi?
- X ve Y değişkenleri doğrusal olmayan bir şekilde ilişkili mi?
- Y'deki değişim X'e bağlı olarak değişir mi?
- Aykırı değerler var mı? [9]

2.3.3.7. Q-Q Grafiği

Q-Q grafiği $x_1, x_2, \dots, x_n$ veri setinin normal dağılıma sahip olup olmadığını belirlemek için kantilleri (quantile) kullanan grafiksel bir yöntemdir [15]. Bir q-q grafiği, ikinci veri kümesinin miktarlarına karşı birinci veri kümesinin miktarlarının grafiğidir. 45 derecelik bir referans çizgisi de çizilir. İki veri seti aynı dağılıma sahip bir popülasyondan geliyorsa, noktalar yaklaşık olarak bu referans çizgisi boyunca düşmelidir. Bu referans çizgisinden ne kadar uzaklaşırsa, iki veri kümesinin farklı dağılımlara sahip popülasyonlardan geldiğine dair kanıtlar o kadar büyük olur.



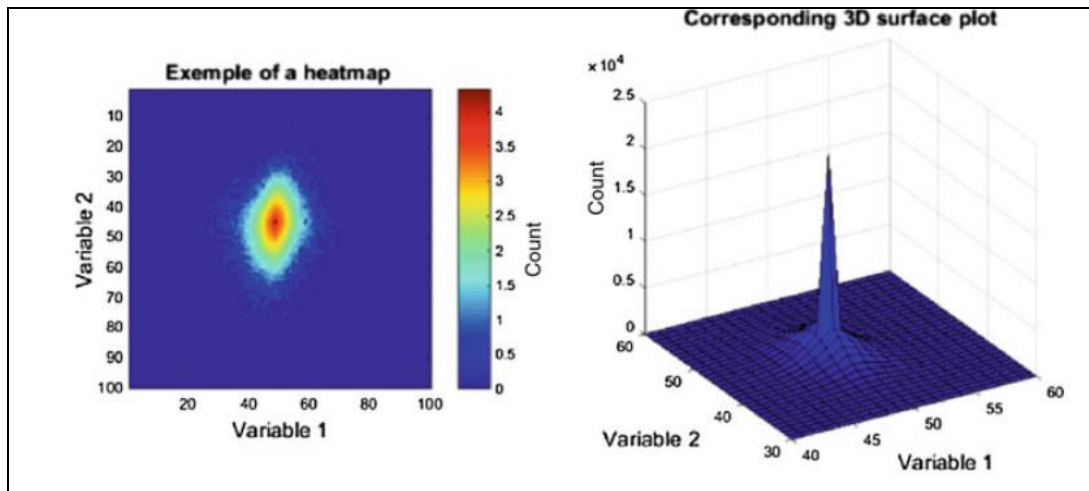
Şekil 2.11: Q-Q grafiği.

Q-Q grafiğinin avantajları:

- Örnek boyutlarının eşit olması gerekmez.
- Dağılımla ilgili olarak bilgi verebilecek birçok argüman aynı anda test edilebilir. Örneğın, konumdaki değışimler, ölçek değışimleri, simetri değışiklikleri ve aykırı değierlerin varlığı bu grafikten tespit edilebilir. Örneğın, iki veri kümesi, dağılımları yalnızca konumdaki bir kayma ile farklılık gösteren popölasyonlardan geliyorsa, noktalar 45 derecelik referans çizgisinden yukarı veya aşağı yer değıştiren düz bir çizgi boyunca uzanmalıdır.
- Ayrıca normallik varsayım testi olarak da kullanılabilen bu grafik oldukça faydalıdır [16][17].

2.3.3.8. Isı Haritaları ve Yüzey Grafikleri

Isı haritaları, rengi her hücrenin değierine bağılı olan 2B diziden oluşturulmuş 2B ızgaralardır. Veri kümesi, hücreleri sonuç değışkeninin değierlerini içeren bir 2B diziye karşılık gelmelidir. Bu teknik, bir sonuç değışkenindeki (ör. kalış süresi) değışikliği diğier iki değışkenin (ör. yaş ve SOFA puanı) bir fonksiyonu olarak göstermek istediğınızde kullanışlıdır. Renk eşlemesi özelleştirilebilir (ör. gökkuşağı veya gri tonlamalı) [12].



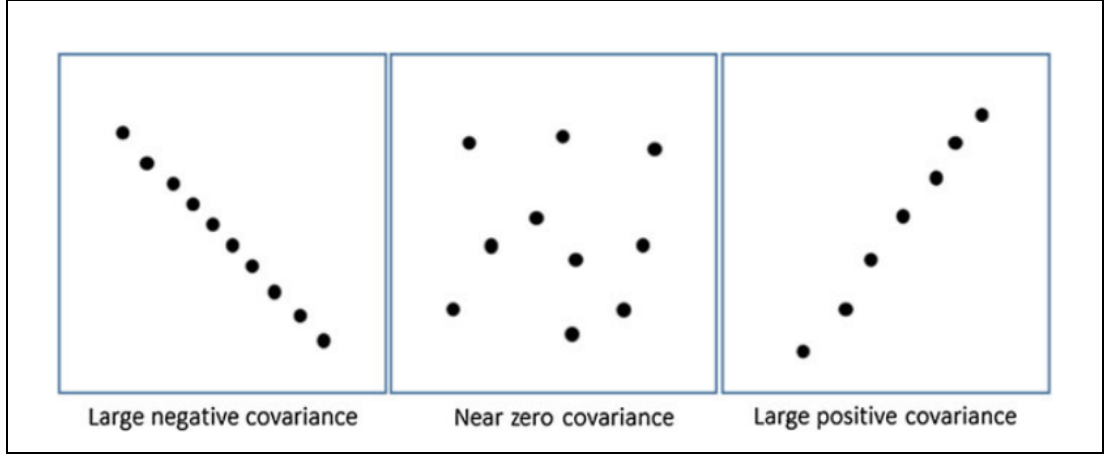
Şekil 2.12: Isı haritaları ve yüzey grafikleri.

2.3.4. Korelasyon ve Kovaryans

Korelasyonlar, değişkenler arasındaki ilişkilerin gücünü tanımlar. Korelasyonları hesaplamanın birden fazla yolu olsa da çoğu elektronik tablo yazılımında yerleşik formül Pearson çarpım momenti korelasyonudur. Bu formül, negatiften pozitif kadar bir aralıkta bir korelasyon skoru üretir ve sıfır hiçbir korelasyonu temsil etmez. Skor uçlara ne kadar yakınsa, ilişki o kadar güçlüdür. Korelasyonlar yalnızca bir ilişkinin gücünü rapor edebilir ve nedensellik hakkında rapor vermez. Uygulayıcıların sonuçları bu gerçeğin ışığında değerlendirmeleri önemlidir (Martin & Bridgmon, 2012).

Yeterince sağlam bir veri kümesine sahip uygulayıcılar, yine çoğu elektronik tablo yazılımında oldukça kolay bir şekilde gerçekleştirilen bir korelasyon matrisinin oluşturulmasında değer bulmalıdır. Bu matrisler, tüm değişkenler arasındaki korelasyonları birkaç basit tıklamayla tek bir grafiğe yerleştirir. Uygulayıcılar, tüm değişkenleri kullanarak bir korelasyon matrisi oluşturmalıdır. Bazı durumlarda, uygulayıcıların bir kodlama işlemi yoluyla dizi değişkenlerini (metin gibi sayısal olmayan değişkenler) sayısal değişkenlere dönüştürmesi gerekebilir [8].

Kovaryans, iki değişkenin ne kadar “ortak değişkenlik gösterdiğinin”, yani bir değişkenin diğeri değiştiğinde ne kadar (ve hangi yönde) değişmesini beklememiz gerektiğinin bir ölçüsüdür. Pozitif bir kovaryans, değişkenlerin pozitif olarak ilişkili olduğu (aynı yönde birlikte hareket ettikleri) anlamına gelirken, negatif bir kovaryans, değişkenlerin ters ilişkili olduğu anlamına gelir. Kovaryansla ilgili bir problem, değerinin rastgele değişkenlerin değerlerinin ölçeğine bağlı olmasıdır. Kovaryansların yorumlanması zor olma eğilimindedir, bu nedenle bunun yerine genellikle korelasyon kullanırız.[10] Korelasyonun her zaman -1 ile +1 arasında olması gibi güzel bir özelliği vardır, -1 süper negatif doğrusal korelasyon, +1 süper pozitif doğrusal korelasyon ve 0 ise X ve Y'nin arasında bir ilişki olmadığını gösterir [12].



Şekil 2.13: Kovaryans örnekleri.

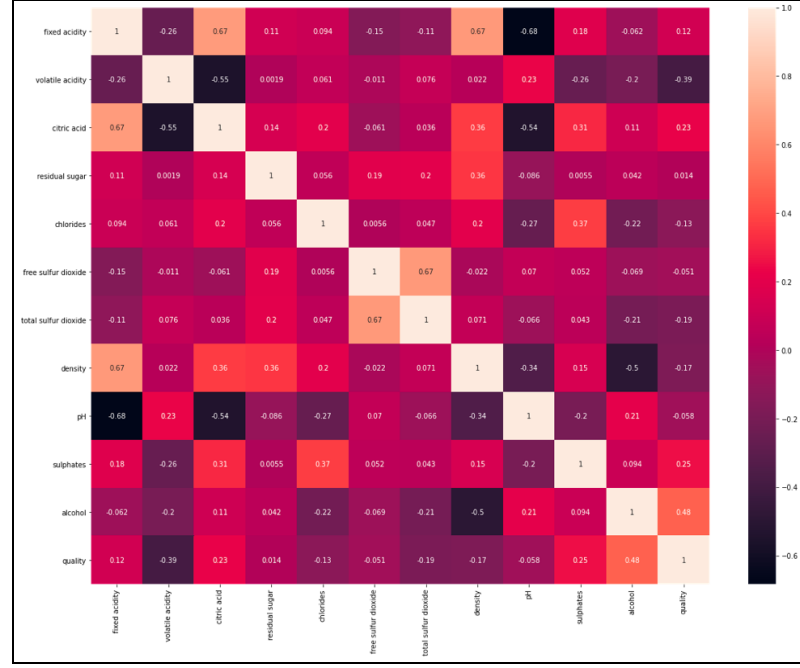
Çok sayıda nicel değişkenimiz olduğunda, en yaygın grafiksel olmayan Keşifsel Veri Analizi tekniği, tüm ikili kovaryansları ve/veya korelasyonları hesaplamak ve bunları bir matriste bir araya getirmektir. İki rastgele değişken arasındaki korelasyon, -1'den 0'a kadar +1 arasında değişen bir sayıdır ve sırasıyla güçlü bir ters ilişkiyi, ilişki olmadığını ve güçlü bir doğrudan ilişkiyi gösterir [18],[19].

Kovaryans için genel formül aşağıdaki gibidir.

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \quad (2.6)$$

Korelasyon için genel formül aşağıdaki gibidir. Burada s_x , X'in standart sapmasını ifade ederken s_y Y'nin standart sapmasını temsil eder.

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{s_x s_y} \quad (2.7)$$



Şekil 2.14: Korelasyon matrisi.

2.3.5. Boyut Analizi

Faktör analizi ve temel bileşen analizi (PCA), muhtemelen ilişkili değişkenlerden oluşan bir korelasyon matrisini, temel bileşenler adı verilen daha az sayıda faktöre dönüştürür. Küme analizi gibi, faktör analizi de boyutları azaltır. Bileşenler, temeldeki (gizli) değişkenleri tanımlar. Faktör analizi ayrıca hangi değişkenlerin “birlikte gittiğini” belirleyebilir. İlk bileşen, verilerdeki değişkenliğin mümkün olduğu kadar çoğunu açıklar ve sonraki her bileşen, kalan değişkenliğin mümkün olduğu kadar çoğunu hesaba katar [20].

2.3.5.1. Temel Bileşen Analizi

Temel Bileşen Analizi (PCA), değişkenleri daha küçük bir boyuta indirgeyerek verileri daha anlaşılır bir hale getirir. Temel Bileşen Analizi, sütun bazında sıfır ortalamalı bir veri matrisi $X = \{x_i | x_i \in \mathbb{R}^D\}$, $d \ll D$ 'nin şu şekilde olduğu düşük boyutlu bir $T = \{t_i | t_i \in \mathbb{R}^d\}$ temsiline dönüştürür:

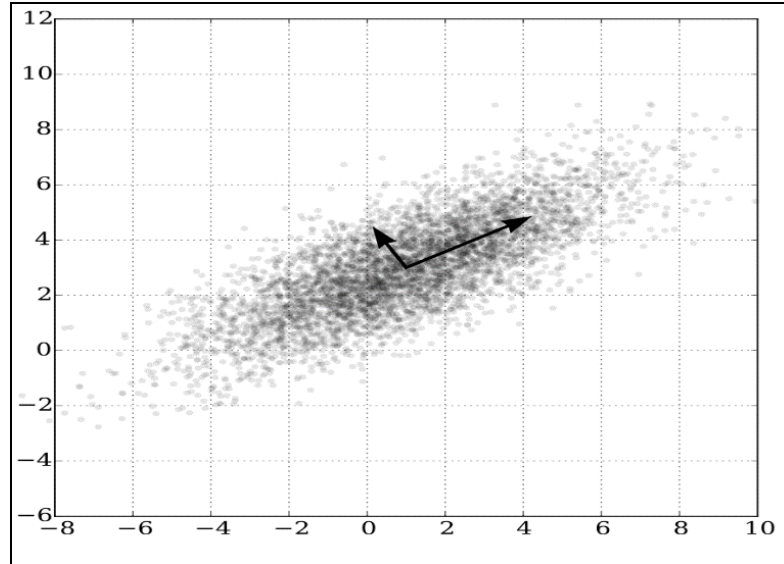
$$T = X \cdot W^T \quad (2.8)$$

W, T'nin X'te maksimum olası varyansı koruduğu şekilde tanımlanan bir $d \times D$ matrisidir. Kavramsal düzeyde Temel Bileşen Analizi, varyansı paylaşan değişken kümelerini tanımlar ve bu varyansı temsil edecek bir komponent yaratır. Sezgisel olarak, PCA'nın yaptığı şey, yüksek boyutlu bir girdi veri kümesini, doğrusal bir dönüşüm yoluyla orijinal verilerin mümkün olduğunca fazla varyansını yakalayan düşük boyutlu bir temsile dönüştürmektir. Bu nedenle, indirgenmiş T temsiliinde, her sütun, orijinal X verisindeki sütunların doğrusal bir kombinasyonu olarak düşünülebilir.

PCA, varsayımlardaki basitliği nedeniyle sıklıkla kullanılan bir yöntemdir. W'nin bulunması genellikle X'in kovaryans matrisinin öz ayrıştırılmasıyla elde edilebilir ve verimli bir şekilde hesaplanabilir.

Adımları şu şekilde özetleyebiliriz.

- Her komponent için ortalama vektör hesaplatılır.
- Kovaryans matrisleri çizdirilir.
- Her komponent için özvektörleri ve özdeğerleri hesaplatılır
- Her bir özdeğerin yüzdesi alınarak varyansın açıklama oranı belirlenir.



Şekil 2.15: PCA Grafik gösterimi.

2.3.6. Kümeleme

Kümeleme analizi veya sınıflandırma, bazı sonlu kümelerden benzer öğelerin (özneler veya değişkenler) kümeleri halinde öğelerin (özneler veya değişkenler) gruplandırılması için bir dizi çok değişkenli yöntemi ifade eder. Farklı küme analizi türleri vardır. En popüler olanları şunlardır: hiyerarşik küme analizi ve K-means kümesi. Hiyerarşik küme analizi, aşamalı adımlarla benzer durumları belirlemeye çalışır. Bir dendogram (popülasyonun ağaç diyagramı) üretilmesine izin verir. Dendogram, vakaların yakınlığını (mesafesini) gösterir [20].

K-Means, kümeleme algoritmalarının içinde en çok kullanılan yöntemlerden biridir. "Unsupervised" dediğimiz etiketsiz öğrenme prensibine sahiptir. Oldukça hızlı çalışmaktadır. İlk kez K-Means ismi 1967 yılında J. B. MacQueen tarafından kullanılmıştır. Olumsuz yönlerinden birisi k küme sayısının kullanıcı tarafından belirlenmesidir. Kullanıcı kendi sezgisel yöntemleri ile optimum k değerini belirlemek zorunda kalır ve bunun için algoritmayı sürekli yeniler. Çünkü ilk yapılan analiz yeterli olmayabilir [21].

K-means kümeleme, m noktayı, her noktanın kümeyle ait olduğu k kümeyle bölmeyi amaçlar, böylece noktalardan atanan küme merkezlerine olan karelerin toplamı en aza indirilir. Verileri karakterize etmek için k adet küme merkezini kullanır.

$$\text{Arg min} \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2, \mu_i = \text{ortalama}(S_i) \quad (2.9)$$

Optimum k seçimi, veri noktalarının dağılımına ve ayrıca istenen kümeleme çözünürlüğüne bağlı olarak genellikle belirsizdir [22].



Şekil 2.16: Kümeleme grafiği.

Basitçe algoritma adımları aşağıdaki gibidir.

- Rassal küme merkezlerini belirle
- Her bir değer ile bu merkez arasındaki uzaklığı hesapla
- Mesafesi en az olan değeri ilgili kümeye ata.
- Küme merkezini tekrardan hesapla
- Uzaklıkları tekrar hesapla
- Eğer atanacak veri noktası kalmadıysa devam et eğer kaldıysa algoritmaya devam et [21].

2.3.7. Değişken Seçimi

Analizlerde değişken veri setimizde kolonlarda bulunan değerlerdir. Bu değerler ile önemli yorumlar ve çıktılar elde ederiz. Değişken seçimi, veri setimiz içerisindeki en önemli olanları seçme işlemidir. Bu işlem model performansını çok fazla etkilemektedir. Değişken seçiminde etiket verilerinin düzeyine bağlı olarak

denetimli, denetimsiz ve yarı denetimli yaklaşımlar kullanılır. Gereksiz değişkenler modelinizin eğitim süresini arttırabilmektedir. Sapmalara neden olabilmekte. Çalışma süresini uzatabilmektedir. Ayrıca fazla sayıda olan değişkeni yorumlamak karmaşıklığa yol açabilir. Değişken seçim metotlarının büyük bir kısmı 3 ana kategoriye ayrılabilir.

2.3.7.1. Filtreleme Yöntemleri

Değişkenlerin önemini hesaplamak için değişken ile hedef değer arasındaki ilişkiyi inceleyen bir yöntemdir. Bu işlem sonucunda veri setini filtrelemiş oluruz ve belirlenen değişkenleri seçerek başka bir alt küme yaratırız. Filtre yöntemleri iki adımdan oluşur: ilk adım değişkenleri tek değişkenli veya çok değişkenli şekilde bazı ölçülere göre sıralar, ikinci adım düşük dereceli özellikleri kaldırır. (Li ve diğerleri. (2018)).

Filtreleme yöntemlerinin avantajları:

- İstenilen modellerde çalıştırılabilir.
- Oldukça hızlı ve zaman harcamayacak şekilde çalışır.
- Çeşitli istatistiksel metotlar ile kullanılması mümkündür.

olarak sayılabilir.

Filtreleme yöntemlerinin dezavantajları:

- Bu yöntem daha önce de bahsettiğimiz gibi hedef değer ile değişken arasındaki ilişkileri incelemektedir. Buradaki soru işareti önemsiz olarak gördüğümüz değişkenlerin topluca etkisinin olumlu olarak maskelenmiş olabileceğidir [24],[25].

olarak sayılabilir. Filtreleme yöntemleri aşağıdaki gibi açıklanabilir.

- Pearson Korelasyon

Doğrusal ilişkiye sahip özellikler arasındaki ilişkileri ölçerken bu yöntemi kullanabiliriz. Bu yöntem literatürde oldukça fazla kullanılmaktadır.

Değişkenlerimizin nümerik değerler olması bu yöntem için önem arz etmektedir. Yapılan test özellikler arasındaki ilişkinin oranını göstermektedir.

- i) Sonuçlar -1 ile +1 arasında hesaplanır.
- ii) 0 veya 0'a yakın değerlere sahip değişkenler arasında korelasyon yoktur.
- iii) 1'e yakın değerler arasında olumlu bir ilişki vardır.
- iv) -1'e yakın değerler arasında olumsuz bir ilişki vardır.

- Spearman Korelasyon

Bu yöntem Pearson yöntemi ile benzerdir. Yani elde edilen sonuçları aynı şekilde değerlendiririz. Burada en önemli nokta hem parametrik hem de parametrik olmayan verilerde kullanılabileceğidir. Ama Pearson için aynı şeyler söylenemez orada verilerin nümerik olması gerekmektedir.

- Kendall Korelasyon:

Burada veriler -1 ve +1 olmak üzere değerler olarak değerlendirilmektedir. Nümerik olmayan veriler için kullanılır. En çok kullanım şekli ise sıralanmış verilerde kullanımıdır.

- ANOVA Yöntemi

ANOVA (Analysis of Variance) yöntemi, değişkenlerdeki varyansları analiz ettiğimiz bir yöntemdir. Bu testi yaparken değişkenler arasında doğrusallık olduğunu varsayırız. Ayrıca normal dağılıma uygunluklarını kabul ederiz. Değişkenlerden birinin kategorik olduğu durumlarda çoğunlukla kullanılmaktadır. Değişkenlere göre anlamlılıklar test edilir. Bu anlamlılık düzeyi P olarak ifade edilir.

- Ki-Kare Testi Yöntemi

Nümerik değerler için çalıştırılması uygun olmayan bu yöntem kategorik değişkenler için çalıştırılır. Ki-Kare testi yöntemi hedef değişken ve öznitelik arasında korelasyon bulunup bulunmadığını belirlememize yardımcı olur. Bunu belirlerken p değerlerini kullanır. Örneğin p değeri 0.05 den küçükse değişkenimizin bağımlı olduğunu söyleyebiliriz. Aksi halde değişkenimiz bağımsızdır.

2.3.7.2. Sarmalayıcı Yöntemler

Bu yöntem model performansını ölçerek. Bir alt özellik grubuna göre model oluşturur ve bu modeller optimum sonucu verene kadar defalarca kurulur ve çalıştırılır. Eğer optimum sonuca ulaşırsa durulur. Bahsettiğimiz gibi bu yöntem modelleme yaparken kullanılır [24],[25].

- İleri Doğru Seçim

Bu yöntemde performansı en yüksek değişken seçilerek aşamaya başlanır. Tek tek diğer değişkenlerde ilk değişken gibi seçilir. Döngü optimum performans ölçütüne ulaşana kadar sürer.

k öznitelikli bir veri setinde:

- i) İlk adımda en yüksek performansa sahip değişken seçilir.
- ii) İkinci adım olarak en iyi tahminlemeye sahip diğer değişken seçilir.

Seçimler kriterlerimize ve ölçütümüze bağlı olarak en performanslı değişkenlerin seçimine kadar devam eder [24],[25].

- Geriye Doğru Seçim

Bu yöntemde en kötü performans gösteren değişkenleri eleyerek en iyi performans gösteren değişkenleri elde etmiş oluruz. Bu yöntem ileriye doğru seçim yönteminin tam tersi olarak düşünülebilir.

k adet öznitelikli bir veri setinde;

- i) İlk adımda en kötü performansa sahip değişken seçilir. Bu değişken kaldırılır.
- ii) İkinci adımda tekrardan en kötü performans gösteren değişken seçilir ve kaldırılır.
- iii) Bu işlemler optimum sonuca ulaşana kadar devam eder [24],[25].

- Adımsal Seçim

Bu yöntem hem ileri hem geriye doğru seçim yöntemlerinin bir karışımı olarak düşünülebilir. Bu uygulamaların birleşiminden oluşur. Burada değişkenleri bütünsel

olarak değerlendiririz. Yani bir değişkeni serideki tüm değişkene göre değerlendirmeye tutarız [24],[25].

2.3.7.3. Gömülü Yöntemler

Burada çeşitli algoritmaların kendine özgü seçim yöntemleri bulunmaktadır. Burada algoritmalar sezgisel olarak seçimlerini yaparlar. Modele alınan değişkenin modeldeki performansı gözlenir. Performansa göre bu etki artırılıp azaltılarak değişkenin modeldeki etkinliği ölçülür.

Gömülü yönteminin avantajları:

- En performanslı değişkenler girdi olarak modele alınır.
- Hızlı çalışmaktadır.
- Doğruluk oranı fazladır.
- Aşırı uyum problemine karşı oldukça hassas davranır [24],[25].

gibidir.

- LASSO (L1)

Doğrusal Regresyonun maliyet fonksiyonuna ek olarak terim eklenir. Hata kareleri toplamını minimize etmeye çalışır. Buradaki önemi az olan değişkenler katsayıları sıfıra indirgenir [24].

- RIGDE (L2)

Lassodan farklı olarak önemi az olan değişkenlerin katsayılarını sıfıra indirmemez. Burada hata kareler toplamını minimize eden katsayıları bulmaktır.

- Elastic Nets (L1/L2)

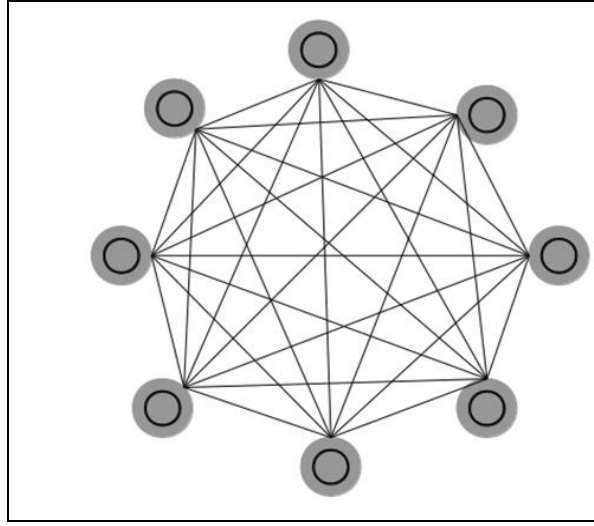
Bu yöntem LASSO ve RIGDE'nin bir yöntemlerinin bir kombinasyonu olarak düşünülebilir. Her iki yöntemde kullanarak daha uygun değişken kümelerini bulmaya yarar [25].

2.3.7.4. Grafiksel Modeller

- Markov Ağları

Markov zinciri geçmiş ve gelecekteki değişkenlerin bağımsız olduğu rastgele değişkenler serisidir. Markov ağların yönlendirilmiş ağlardır. Bu seri, $X_1, X_2, X_3, \dots$ ile ifade edilir. Her X değişkeni için S değerleri zincirin durumunu belirten bir seri oluşturur. Yönlendirilmiş doğrulardaki olasılıklar aşağıdaki gibidir. Kenarlarda bulunan iki durum arasındaki geçiş olasılık ile tanımlanır [25].

$$\Pr(X_{n+1} = x | X_n = x_n, \dots, X_1 = x_1) = \Pr(X_{n+1} = x | X_n = x_n) \quad (2.10)$$



Şekil 2.17: Markov zinciri.

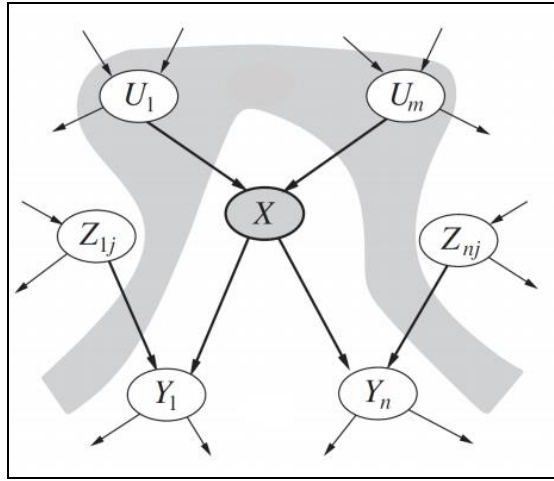
- Bayes Ağları

Bayes teoremi rassal olaylar arasındaki (Örneğin A ve B) koşullu olasılıkları ve marjinal olasılıklar arasındaki ilişkileri temsil eder.

İlgili terimler aşağıdaki gibidir.

- $P(A)$ = A için marjinal olasılık. Burada A, B ile ilgili bir bilgiye sahip değildir.
- $P(A|B)$ = B için, A'nın koşullu olasılığı.
- $P(B|A)$ = A için, B'nin koşullu olasılığı.
- $P(B)$ = B'nin marjinal olasılığı

Kısacası Bayes ağları yönlendirilmiş ağlardır ve B gözlemine göre A gözleminin değişimlerini inceler [26],[27].



Şekil 2.18: Bayes teoremi.

- Bağımlılık Ağları

Bağımlılık ağları (Dependency Network-DN), Markov ağlarına benzer grafik modellerdir, burada her bir köşe (düğüm) rastgele bir değişkene karşılık gelir ve her kenar değişkenler arasındaki bağımlılıkları yakalar. Bayes ağlarının aksine, DN'ler döngüler içerebilir. Her düğüm, ebeveynleri verilen rastgele değişkenin gerçekleşmesini belirleyen koşullu bir olasılık tablosuyla ilişkilendirilir [28]. Ortak dağılım $p(x)$ ile $X=(X_1, X_2, X_3, \dots, X_n)$ rastgele değişkenler kümesi için tutarlı bir bağımlılık ağı bir çifttir (G,P) burada G döngüsel yönlendirilmiş bir grafiktir, burada düğümlerinin her biri X deki bir değişkene karşılık gelir ve P bir dizi koşullu olasılık dağılımıdır. P_{xi} ile gösterilen X_i düğümünün ebeveynleri, aşağıdaki bağımsızlık ilişkilerini karşılayan $P_{xi} \subseteq (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ değişkenlerine karşılık gelir [29].

$$p(x_i | p_{xi}) = p(x_i | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = p(x - x_i) \quad (2.11)$$

2.3.7.5. DiscoVars Yaklaşımı

Bu yaklaşım, modelleyicinin bağımlılık ağlarını ve grafik merkezilik ölçülerini kullanarak etkileşimli bir yöntemle aday değişken alt kümesini keşfetmesini sağlar.

Denetimli öznitelik seçim yöntemleri somut performans ölçüleri üzerine kurulduğundan, bu çalışmada kümeleme analizi için öznitelik seçim yönteminin uygulanması tercih edilmiştir. Bu yaklaşımdaki yenilik, adımsal, ileri ve LASSO gibi değişken seçim yöntemleriyle oluşturulan bağımlılık ağlarındaki (DN) değişkenlerin önemini belirlemek için grafik merkezilik ölçümlerini kullanmaktır [30].

Kısaca adımlar şu şekildedir.

- DN Grafik modelin oluşturulması
- Oluşturulan grafik model kullanılarak hesaplanan merkezilik ölçümüne dayalı olarak değişkenlerin sıralanması
- Önemli değişkenleri sıralamak için M değişken yönteminin kullanılması
- Top-n değişkenler ile birlikte kümeleme analizinin yapılması

Yaklaşımda temel olarak, tüm değişkenler arasındaki tüm anlamlı ilişkileri temsil etmek için her X_i değişkenini kalan $X \setminus X_i$ değişkeniyle eşleştirerek bir bağımlılık ağı (DN) oluşturmaya dayanır. Yerel dağılımlardaki önemli değişkenleri belirlemek için istatistiksel değişken seçim yöntemi M kullanılmalıdır. Varsayımlara göre s kümesi X_i için anlamlı değişkenlerin endekslerini temsil eder ve X_s değişkenleri X_i nin ebeveyn Pai dir. Yani X_i, X_s değişkenlerine bağlıdır. Bu ilişkilerin tümü DN G 'yi oluşturur. Adımsal, İleri Doğru Seçim, AIC ve Lasso seçim yöntemleri, bu yaklaşımda lineer regresyon modelleri (lm) tarafından kullanılabilir. Discoverys Yaklaşımı için geliştirilmiş olan algoritma aşağıdaki gibidir.

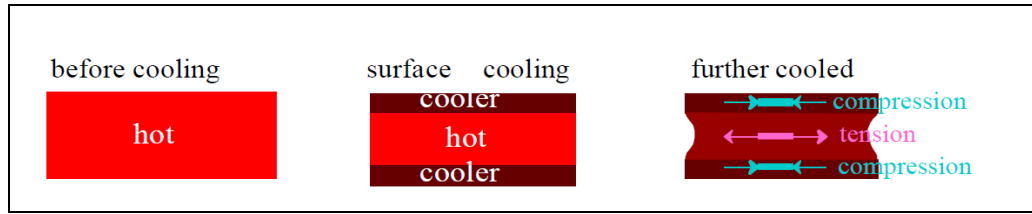
Tablo 2.2: DiscoVars algoritması.

Algorithm 1 DiscoVars: Construct DN G
1: Given X and M
2: for $i \leftarrow 1, d$ do
3: Fit $X_i \leftarrow \text{lm}(X \setminus X_i, M)$
4: $s \leftarrow \{\text{significant variables}\}$
5: $E_{is} \leftarrow 1$
6: end for
7: $c \leftarrow \text{Select Centrality Measure}$
8: Rank(G, c, X)
9: Select Top – n variables

Algoritma seçim yöntemi M , merkezilik ölçüsü c ve İlk n önemli değişkenler için n seçimlerine dayanmaktadır. Bu seçimler sonucunda bize önemli değişkenleri sunar.

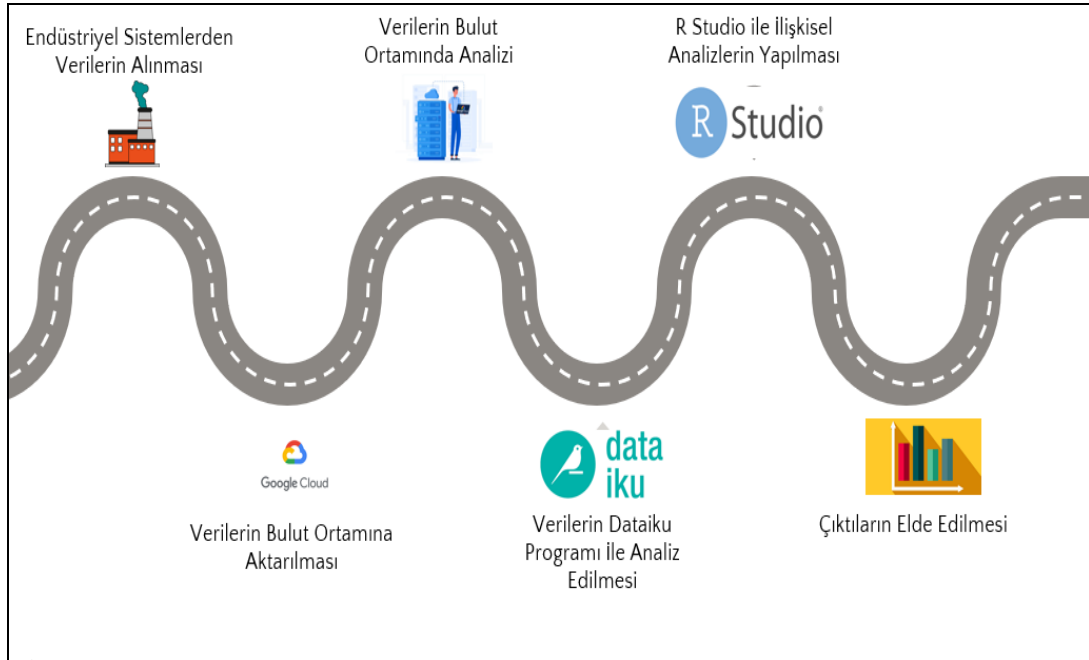
3.SAYISAL SONUÇLAR ve YORUMLAR

Bu çalışmada, cam üretim tesislerinde kullanılan ısı sensörlerindeki veriler ele alınmıştır. Örneğin; dönüş su sıcaklığı, boyun su sıcaklığı gibi. Bu sensörler camın aşama aşama soğutulduğu prosesteki soğutma suyu sıcaklıklarını ifade eder. Cam normal olarak soğutmaya bırakılırsa gerilim oluşacağı için kırılmaya uğrayabilir. Bu gerilimin giderilmesi için kontrollü bir şekilde soğutulması gerekir. Bu değerler ekranlarda 7/24 canlı takip edilmektedir ve sensör değerlerine göre ilgili aksiyonlar alınmaktadır.



Şekil 3.1: Cam soğutma gösterimi.

Çalışmanın genel akışı aşağıdaki gibidir.



Şekil 3.2: Veriyi elde etme ve analiz yol haritası.

3.1. Analizde Kullanılan Programlar

3.1.1. Google Cloud Platform (GCP)

Google Cloud, sunucu altyapı hizmetlerini son kullanıcıya sunduğu bir bulut bilgi platformudur [31]. Google Cloud Console ile SQL dili kullanılarak veri dönüştürme veya manipülasyon gibi analizler yapılabilir.

3.1.2. Dataiku

Dataiku, 2013 yılında kurulmuş bir yapay zeka ve makine öğrenimi şirkettir. Bu platformda çok sayıda makine öğrenmesi algoritması bulunmaktadır. Bu algoritmalar açık kaynaklı olarak sunulmuştur.

3.1.3. RStudio

RStudio, istatistiksel analizler ve grafikler için kullanılan R programlama dili ile entegre çalışır. RStudio da kullanılan R dili, istatistiksel hesaplamalar yapmak veya grafikler çizdirmek için kullanılan bir yazılım ortamıdır. R Foundation tarafından desteklenen özgür bir yazılımdır. Yeni Zelanda Auckland Üniversitesinden Ross Ihaka ve Robert Gentleman tarafından geliştirilen R, S programlama dilinin uyarlaması olarak değerlendirilebilir [32].

R istatistikçiler tarafından standart olarak veri analizi alanında kullanılmaktadır. Bu analizler özelinde çok sayıda kütüphaneye ve açık kaynak koduna sahiptir. RStudio da genel olarak komut satırları bulunsun da kullanıcıya sunulabilecek çeşitli grafiksel yüzlere de sahiptir [33].

3.2. Veri Kümesinin Detayları

Veri Kümesi GCP (Google Cloud Platform)'da bulunmaktadır. Veri Kümesinde 23 adet değişken bulunmaktadır. Bu değişkenlerden 20 tanesi sensör verilerini ifade etmektedir. Bu veriler double formattadır. Diğer 3 değişken ise tarih ve nümerik formatta değişkenlerdir.

Tablo 3.1: Veri detayları.

SENSOR	SENSOR ADI
SENSOR600	Bel Mikseri Çıkışı Soğutma Suyu Sıcaklığı 1
SENSOR601	Bel Mikseri Çıkışı Soğutma Suyu Sıcaklığı 2
SENSOR602	Bel Mikseri Çıkışı Soğutma Suyu Sıcaklığı 3
SENSOR603	Bel Mikseri Çıkışı Soğutma Suyu Sıcaklığı 4
SENSOR604	Bel Mikseri Çıkışı Soğutma Suyu Sıcaklığı 5
SENSOR605	Bel Mikseri Çıkışı Soğutma Suyu Sıcaklığı 6
SENSOR610	Boyun Soğutucu Su Sıcaklığı Giriş Sol
SENSOR611	Boyun Soğutucu Su Sıcaklığı Çıkışı Sol
SENSOR612	Boyun Soğutucu Su Sıcaklığı Giriş Sağ
SENSOR613	Boyun Soğutucu Su Sıcaklığı Çıkışı Sağ
SENSOR631	Fıskiye Su Sıcaklığı 1
SENSOR633	Fıskiye Su Sıcaklığı 3
SENSOR634	Fıskiye Su Sıcaklığı 4
SENSOR635	Fıskiye Su Sıcaklığı 5
SENSOR636	Fıskiye Su Sıcaklığı 6
SENSOR637	Fıskiye Su Sıcaklığı 7
SENSOR638	Fıskiye Su Sıcaklığı 8
SENSOR639	Fıskiye Su Sıcaklığı 9
SENSOR640	Fıskiye Su Sıcaklığı 10
SENSOR641	Fıskiye Su Sıcaklığı 11

3.3. Veri Kümesi Dönüşümleri ve Standardizasyon

Veri dönüşümünün amacı, veri kümesindeki aykırı değerlerin veya eksik değerlerin olumsuz etkilerini minimuma indirerek performansı ve kaliteyi iyileştirmektir. Verilerimizi Keşifsel Veri Analizi için hazırlamamız gerekmektedir. Verilerimiz GCP bulut ortamında tutulduğu için veri standardizasyon adımlarından birçoğu GCP'nin konsol ekranında düzenlenmiştir. Düzenlenen bu veriler csv formatında dışarı aktarılarak Dataiku ve RStudio programlarında içeri aktarılarak analizlerde kullanılmıştır.

3.3.1. Veri Kaynağından Verilerin Çekilmesi

GCP'de bulunan veri kaynağında bu çalışmada kullanılmayacak birden fazla veri bulunmaktadır. Bu veriler içerisinde ilgili sensör ve tarih verilerimiz

filtrelenerek çekilmiştir. Çekilen bu veriler kullanılarak analizlerimizde kaynak olması açısından yeni bir tablo oluşturulmuştur.

3.3.2. Veri Frekansının Düzenlenmesi

Sistemler sensör verilerini 1 saniyelik frekanslar ile çekmektedir. Veri setinde bulunan olası kaymaları önlemek ve verileri standart hale getirmek için verilerimizi 10 dakikalık frekansa göre ortalaması alınmıştır.

3.3.3. Veri Formatının Düzenlenmesi

Veri setindeki sensör verilerinin formatları düzenlenerek virgülden sonra 4 basamak şeklinde standardize edilmiştir. Bunun nedeni her sensör için basamak uzunluğunun standart olmayıp analizlerde sorun yaratmasıdır. Tarih formatları ise 10 dakikalık ortalamalara göre yeniden düzenlenmiştir. Veriler artan şekilde sıralanmıştır. CONCAT fonksiyonu ile tarih alanı yeniden düzenlenmiştir.

3.3.4. Eksik Verilerin Hariç Tutulması

Veri setinde bulunan bazı sensörlerde tekrar eden veriler veya eksik bulunan verilerin olduğu sensörler bu çalışmada hariç tutulmuştur. Hariç tutulan sensörlerin dahil edildiği çalışmalarda daha az kalitede çıktı üretildiği ve sonuçlarda sapmaya neden olduğu için bu sensörler çalışmadan çıkarılmıştır. İlerleyen zamanlarda bu sensörlerde bulunan eksik veriler doldurularak çalışmaya dahil edilebilir.

3.4. Veri Özet Tablosunun Oluşturulması

Bu çalışmada ilk aşamada Keşifsel Veri Analizi yapılarak veriler ön incelemeye tutulmuştur. Verilerin özet tabloları oluşturularak Min, Max, Median, Mean değerleri incelenmiştir. Buradaki amacımız analiz etmek istediğimiz veriyi anlamaktır.

```
> summary(Veri_Seti_v2$SENSOR600)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  31.52  42.69  46.06  44.47  46.39  49.17
```

Şekil 3.3: Örnek kod ekran görüntüsü.

Hesaplanan değerler Tablo 4.1’de gösterilmiştir.

Tablo 3.2: Veri özet tablosu.

SENSOR	Min	1st Qu.	Median	Mean	3rd Qu.	Max
SENSOR600	31.52	42.69	46.06	44.47	46.39	49.17
SENSOR601	34.37	44.59	46.72	45.91	47.47	50.59
SENSOR602	31.13	42.36	45.95	44.26	46.33	49.15
SENSOR603	35.11	45.61	47.59	46.86	48.38	51.42
SENSOR604	30.49	41.83	45.35	43.69	45.78	48.69
SENSOR605	30.80	42.12	45.77	44.08	46.25	49.24
SENSOR610	23.08	34.06	37.93	36.21	38.68	41.80
SENSOR611	34.70	44.74	47.19	46.16	47.76	50.56
SENSOR612	23.59	34.61	38.45	36.70	39.08	42.05
SENSOR613	33.12	43.28	46.22	44.85	46.50	49.29
SENSOR631	60.78	74.72	77.81	77.89	81.08	96.67
SENSOR633	70.42	87.72	91.43	90.72	94.23	100.11
SENSOR634	72.76	87.27	90.25	89.96	92.95	99.88
SENSOR635	73.03	87.19	90.33	89.99	93.47	99.91
SENSOR636	72.73	86.35	89.54	89.31	92.88	99.96
SENSOR637	76.13	90.28	93.58	92.99	96.98	99.90
SENSOR638	71.53	91.92	94.91	94.16	92.88	99.96
SENSOR639	79.27	91.91	95.27	94.65	98.51	100.04
SENSOR640	75.79	89.89	93.34	92.68	96.39	100.02
SENSOR641	72.53	86.54	90.40	89.91	93.78	99.96

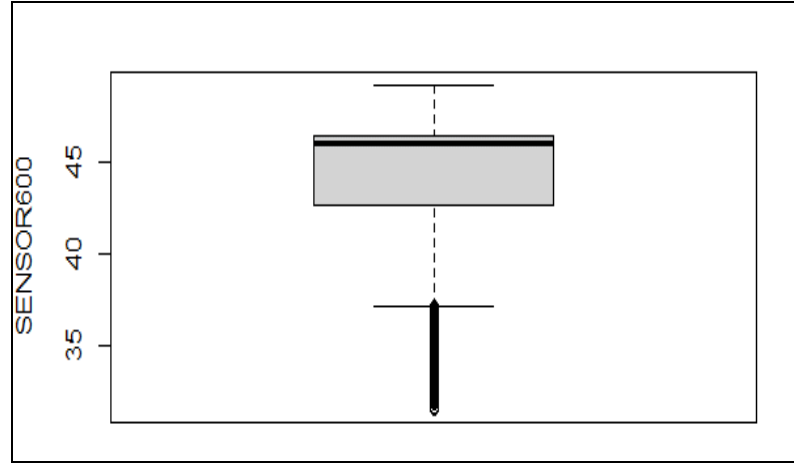
3.5. Aykırı Değerlerin Tespit Edilmesi

Aykırı değerleri tespit ederken RStudio da bulunan aşağıdaki örnek kod çalıştırılmıştır.

```
boxplot.stats()$out
```

Şekil 3.4: Kutu grafiği aykırı verileri hesaplama kodu.

Burada hesaplama yaparken %25 çeyrek değerinden 1,5 kat az ve %75 çeyrek değerinden 1.5 kat fazla olan değerleri aykırı değer olarak sınıflandırdık. Örnek olarak SENSOR600 değişkeni için çizdirilen kutu grafiği aşağıdaki gibidir. Burada Kutu Grafiğimizin altında kalan değerler aykırı verileri temsil etmektedir. 361 adet aykırı değer tespit edilmiştir. Burada aykırı değerlerin toplam veri sayısına oranına bakarsak yaklaşık %0.02 gibi bir oran olduğu için göz ardı edebiliriz.



Şekil 3.5: SENSOR600 kutu grafiği.

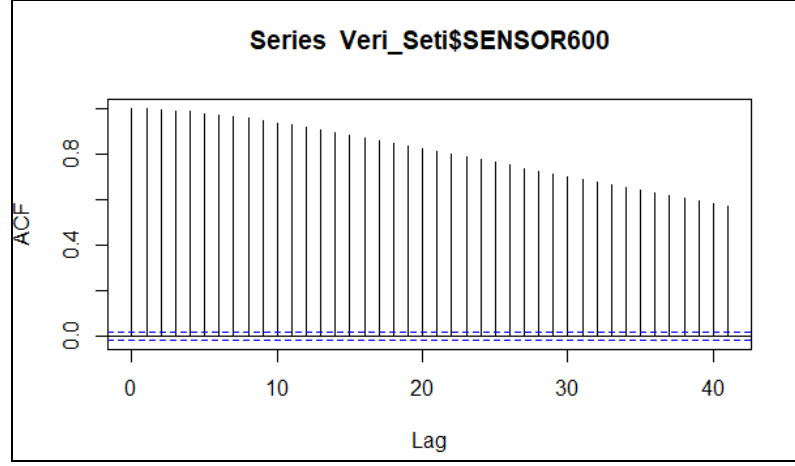
Diğer sensörler ile ilgili kutu grafiği çizimleri **EK-A**'da verilmiştir.

Tablo 3.3: Sensör verilerindeki aykırı değerler.

SENSOR600	SENSOR601	SENSOR602	SENSOR603	SENSOR604
631	588	344	647	342
SENSOR605	SENSOR610	SENSOR611	SENSOR612	SENSOR613
328	260	507	283	436
SENSOR631	SENSOR633	SENSOR634	SENSOR635	SENSOR636
539	324	343	310	303
SENSOR637	SENSOR638	SENSOR639	SENSOR640	SENSOR641
343	350	211	320	289

3.6. Otokorelasyonun Hesaplanması

Sensörde otokorelasyon bulunup bulunmadığı RStudio kullanılarak tespit edilmiştir. SENSOR600 için çizdirilen Otokorelasyon grafiği aşağıdaki gibidir.

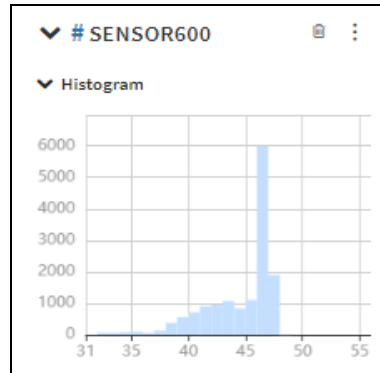


Şekil 3.6: SENSOR600 otokorelasyon grafiği.

Diğer sensörlere ait grafikler **EK-A**'da verilmiştir. Şekil 3.6'yı incelediğimizde verilerin orta düzeyde pozitif otokorelasyona sahip temel bir otoregresif modelden gelmektedir. Grafik LAG 1'de orta derecede yüksek bir Otokorelasyon ile başlamıştır. Doğrusal olarak ilerlemektedir. Uygun bir şekilde modellendiğinde orta düzeyde tahmin edilebilirlik sağlanabilir. Otoregresif model ile parametreler tahmine edilebilir. Diğer sensörlerdeki Otokorelasyon grafikleri incelendiğinde hepsinde pozitif otokorelasyona rastlanmıştır.

3.7. Veri Dağılımlarının İncelenmesi

Histogramlar vasıtasıyla verideki dağılımlar incelenmiştir. Örnek olarak SENSOR600'ün Histogram grafiğini aşağıda gösterildiği gibidir.

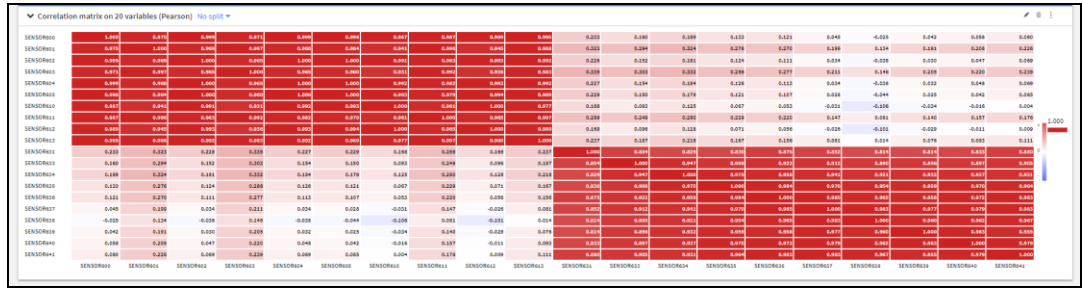


Şekil 3.7: SENSOR600 histogram.

Buradaki dağılımda verilerin sola çarpık olduğu izlenmiştir. Çarpık dağılımlar tahminleme modellerinde karmaşıklık getirir. Bu tip çarpık veriler genellikle alt üst sınırlar nedeniyle oluşur. Alt sınırı olan veriler genellikle sağa çarpıkken, üst sınırı olan veriler genellikle sola çarpıktır. Bu durumda en iyi uyan dağılımı bulup verileri normalize etmek önerilmektedir. Genel olarak dağılımların çarpık olduğu ve normal dağılıma uymadığı görülmüştür. Sensörlere ait diğer histogramlar EK-A’da verilmiştir.

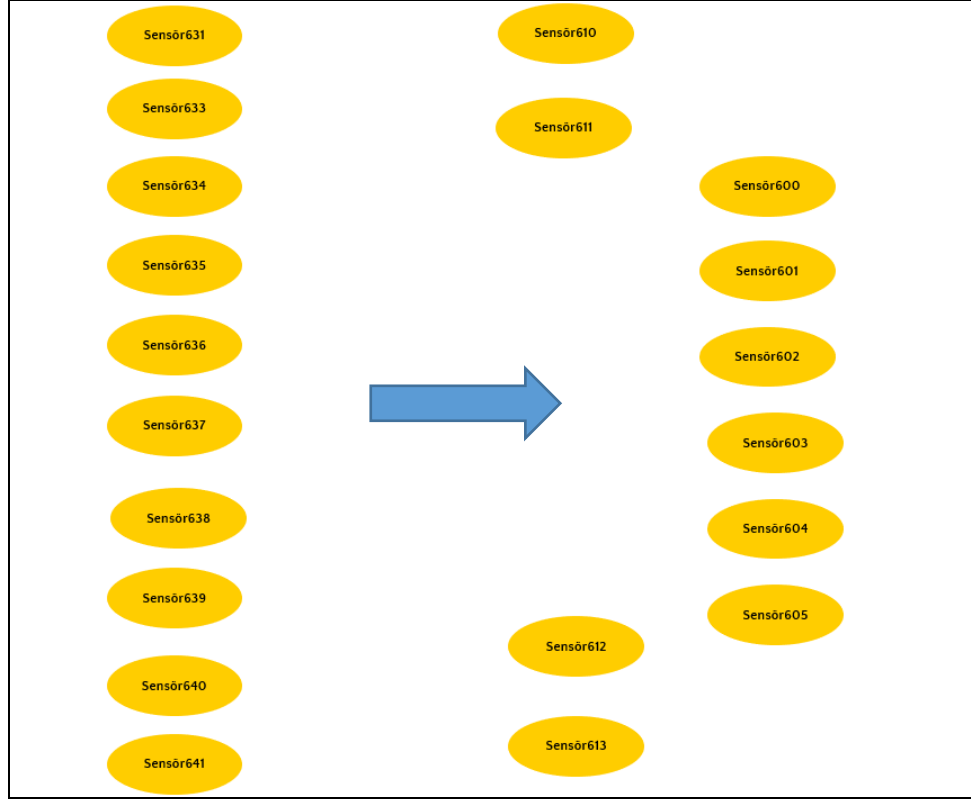
3.8. Korelasyon Matrisi

Sensörler arasındaki ilişkileri tespit etmek için korelasyon matrisi çizdirilmiştir.



Şekil 3.8: Korelasyon matrisi.

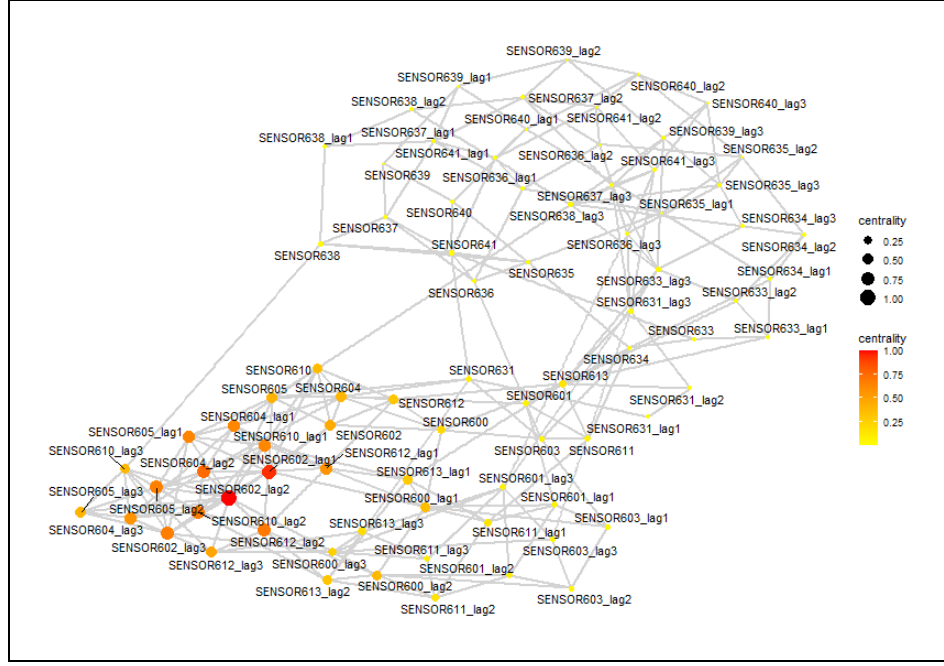
Elde ettiğimiz verilerde aynı bölgede konumlandırılmış sensörlerin birbiri ile pozitif korelasyona sahip olduğu görülmektedir. Farklı bölgelerde konumlandırılmış sensörler arasında da pozitif korelasyon bulunmasına bu ilişki oldukça zayıftır. Örnek olarak SENSOR601 ve SENSOR602 arasındaki korelasyon değeri 1’e çok yakındır. Şekil 3.9 da görebileceğimiz üzere konumlarının birbirine yakınlığı buda ölçüt olabilir. Fakat SENSOR631 ile SENSOR603’ün korelasyon değeri 0.339 görüldüğünden buradaki etkininde araştırması gerekmektedir.



Şekil 3.9: Sensörlerin tahmini konumları.

3.9. Değişken Seçimi ve Kümeleme Algoritmasının Kullanılması

DiscoVars algoritması kullanarak değişkenler incelenmiştir. Değişkenleri incelerken lag 1, lag2 ve lag3 değerleri de veri setine eklenmiş olup gecikmelerde göz önünde bulundurularak analizler yapılmıştır. İlk olarak grafiksel bir model oluşturulmuştur. Merkezi ölçütler kullanılarak değişkenlerin önemi belirlenmiştir. LASSO ve ölçüt olarak eigen kullanarak çizdirdiğimiz grafiksel model aşağıdaki gibidir.



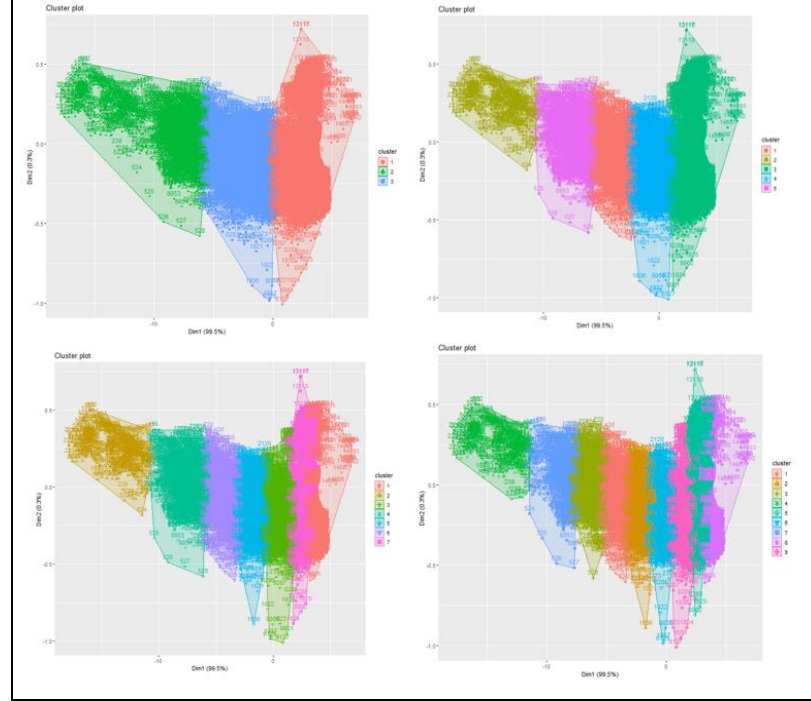
Şekil 3.10: Lasso ve Eigen Value kullanılarak geliştirilen grafiksel model.

Merkezi ölçüt olarak eigen value kullanarak bizim için önemli olan değişkenleri saptarız. Değişkenler aşağıdaki tablodaki gibi bulunmuştur.

Tablo 3.4: Lasso ve Eigen Value kullanılarak bulunan değişkenler.

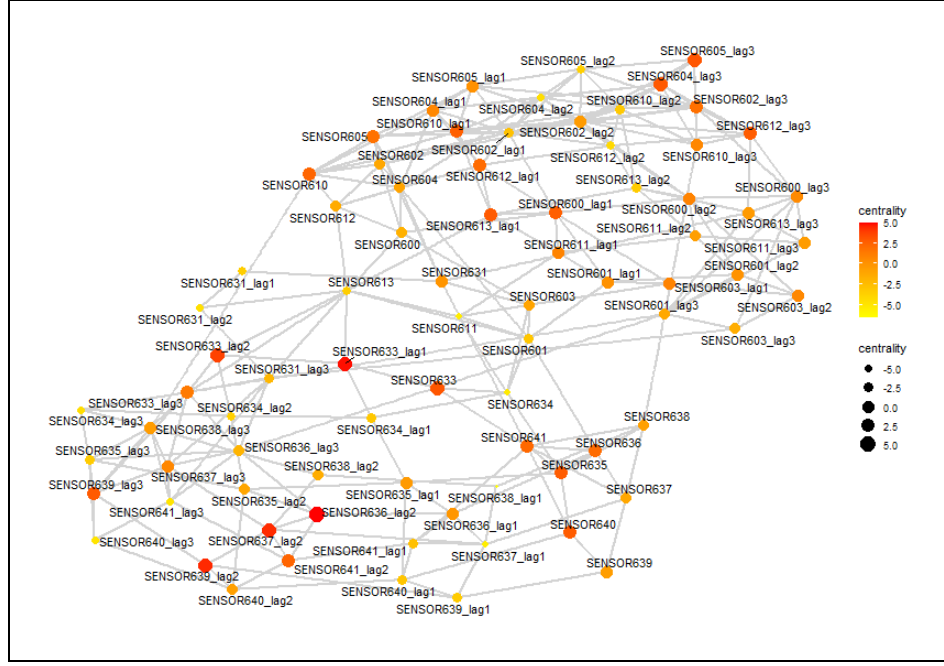
1	SENSOR602_lag2	1
2	SENSOR602_lag1	0.94
3	SENSOR604_lag2	0.75
4	SENSOR610_lag2	0.70
5	SENSOR612_lag2	0.70
6	SENSOR602_lag3	0.68
7	SENSOR605_lag2	0.66
8	SENSOR605_lag1	0.64
9	SENSOR604_lag1	0.63
10	SENSOR610_lag1	0.63
11	SENSOR612_lag1	0.57
12	SENSOR604_lag3	0.56
13	SENSOR612_lag3	0.48
14	SENSOR602	0.45
15	SENSOR605	0.41
16	SENSOR605_lag3	0.40
17	SENSOR604	0.39
18	SENSOR610_lag3	0.39

Bir sonraki aşamada bu değişkenleri kullanarak kümeleme algoritması çalıştırmaktır. Algoritma olarak k-means kullanılmıştır. K değerleri 3,5,7 ve 9 olmak üzere analizler yapılmıştır.



Şekil 3.11: K-Means kümeleme algoritması.

Burada hangi k değerini seçeceğimiz analistin kararına kalmıştır. K=5 im daha iyi bir kümeleme yaptığını görüyoruz. Aynı şekilde analizlerimizi alpha merkezi ölçütüne göre tekrarlıyoruz.



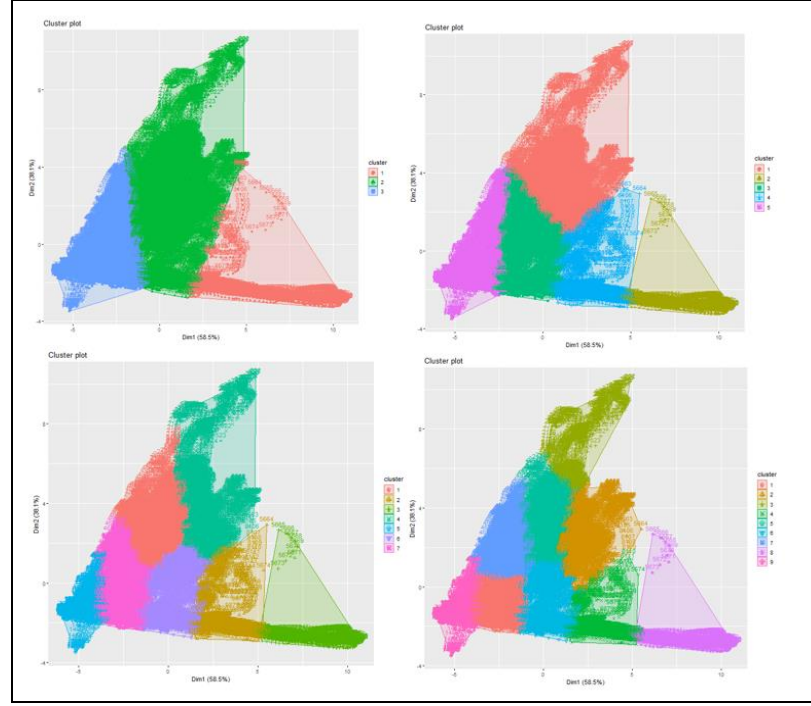
Şekil 3.12 : Lasso ve Alpha ölçütü kullanılarak geliştirilen grafiksel model.

Önemli olan sensörleri aşağıdaki gibi buluruz.

Tablo 3.5: Lasso ve Alpha ölçütü kullanılarak bulunan değişkenler.

1	SENSOR636_lag2	5.019
2	SENSOR633_lag1	4.914
3	SENSOR639_lag2	4.509
4	SENSOR637_lag2	4.464
5	SENSOR633_lag2	3.836
6	SENSOR605_lag3	3.146
7	SENSOR633	3.095
8	SENSOR604_lag3	2.968
9	SENSOR639_lag3	2.920
10	SENSOR613_lag1	2.918
11	SENSOR612_lag3	2.855
12	SENSOR635	2.833
13	SENSOR640	2.827
14	SENSOR610_lag1	2.728
15	SENSOR600_lag1	2.715
16	SENSOR641_lag2	2.483
17	SENSOR610	2.461
18	SENSOR641	2.323

K değerleri 3,5,7 ve 9 olmak üzere analizleri yaparız.



Şekil 3.13: K-Means kümeleme algoritması.

Burada da $k=5$ in kümelemede güzel bir ayrışma sağladığını görebiliriz. Değişkenleri belirlemenin ana fiki tüm sensörleri kullanmak yerine önem dercesine göre seçtiğimiz 18 sensörü kullanarak bir analitik model geliştirilebileceğidir. Ayrıca önemi yüksek olan sensör listesinde bulunan lag'ler sensörlerin gecikmelerden etkilenebileceğini belirtmektedir.

4.SONUÇ

Sonuç olarak, keşifsel veri analizi yaparak verilerimizdeki fenomenleri gözlemledik. Böylelikle veriler hakkında ön gözlem yapma şansına sahip olduk. Verilerimizin dağılımlarını inceleyerek ileride kullanılabilecek modellerde ne gibi bir yol izlemeliyiz veya nelere dikkat etmeliyiz gibi sorulara yanıtlar aradık. Ayrıca yüksek boyutlu veriler ile değişken seçimi yaparken LASSO'nun yüksek performans gösterdiğini gözlemledik. Farklı merkezi ölçütler ile birlikte çizdirilen grafiksel modellerde önem sırasına göre değişkenlerimizi listeledik. Bu değişkenleri ve 10 dakikalık gecikmelerinide dahil ederek k-means algoritması ile kümeleme algoritmasında açıkladık. Kümelerin güzel bir şekilde ayrıştığını tespit ederek yüksek maliyet ve zaman kısıtından etkilenmemek için verilerimizi indirgeyerek daha az veri ile performanslı çıktılar elde edebileceğimizi görmüş olduk. Yüksek hacimli veriler ile çalışırken her bir sensör bizim için önemli olmayabilir. Zaman ve maliyet açısından fayda sağlaması amacıyla önem derecesi yüksek sensörlere odaklanıp analizleri bu yönde devam ettirmek faydalı olacaktır.

KAYNAKLAR

- [1] Web 1, (2022), <https://www.pwc.com/gx/en/technology/pdf/industrial-internet-of-things.pdf>, (Eriřim Tarihi: 21/06/2022).
- [2] Web 2, (2022), <https://www.ictmedia.com.tr/News/Index/15297>, (Eriřim Tarihi: 20/06/2022).
- [3] Tukey John W., (1991), "Data Analysis", Princeton University.
- [4] Behrens J. T., (1997), "Principles and procedures of exploratory data analysis, Psychological Methods", 131-160.
- [5] Web 3, (2002), <https://www.itl.nist.gov/div898/handbook/toolaids/pff/E-Handbook.pdf>, (Eriřim Tarihi: 01/06/2022).
- [6] Anderson R., Jafar M., (2022), "Visual Tools for Initial Exploratory Data Mining".
- [7] Bezerra A., Silva I., Guedes L. A., Silva D., Leitao G., Saito K., (2019), "Extracting Value from Industrial Alarms and Events: A Data-Driven Approach Based on Exploratory Data Analysis".
- [8] Matthew B. C., (2021), "Exploratory Data Analysis in Schools: A Logic Model to Guide Implementation", International Journal of Education Policy & Leadership, 17 (4).
- [9] Churakova A., (2017), "Automating Exploratory Data Analysis for use in Predictive Modeling (Classification)", Yüksek Lisans Tezi, Humboldt Universität zu Berlin.
- [10] Web 4, (2022), <https://r4ds.had.co.nz/exploratory-data-analysis.html>, (Eriřim Tarihi: 10/5/2022).
- [11] Howard J. S., (2018), "Experimental Design and Analysis", Carnegie Mellon University.
- [12] Komorowski M., Marshall D., Saliccioli J., Crutain Y., (2016), "Exploratory Data Analysis", Secondary Analysis of Electronic Health Records, 185-203.
- [13] Web 5, (2021), <https://www.ferdio.com/en/datavizproject>, (Eriřim Tarihi : 05/05/2022) .
- [14] Web 6, (2013), <https://datavizcatalogue.com/TR/yontemleri>, (Eriřim Tarihi : 05/06/2022).
- [15] Web 7, (2020), <https://acikders.ankara.edu.tr/course/view.php?id=6881>, (Eriřim Tarihi: 01/06/2022).

- [16] Chambers M. J., Cleveland W. S., Kleiner B., Tukey P. A., (1983), “Graphical Methods for Data Analysis”, Chapman and Hall/CRC.
- [17] Tukey J., (1977), “Exploratory Data Analysis”, Addison-Wesley.
- [18] Liu Q., (2019), “An Application of Exploratory Data Analysis in Auditing”.
- [19] Web 8, (2020), <https://medium.com/software-development-turkey/machine-learning-2-korelasyon-matrisi-ozellik-seçimi-sınıfların-dengesizliği-karar-aracıları-af993bd8ea66>, (Erişim Tarihi: 01/05/2022).
- [20] Web 9, (2016), https://edutechwiki.unige.ch/en/Methodology_tutorial_exploratory_data_analysis, (Erişim Tarihi: 04/03/2022).
- [21] Takaoğlu M., Takaoğlu F., (2019), “K-Means ve Hiyerarşik Kümeleme Algoritmanın Weka Ve Matlab Platformlarında Karşılaştırılması”, İstanbul Aydın Üniversitesi Dergisi, 11 (3), 303-317.
- [22] Liu K., (2015), “Exploratory Data Analysis on Multivariate Data”, Yüksek Lisans Tezi, Universiteit Leiden.
- [23] Web 10, (2019), <https://medium.com/@gulcanogundur/öznitelik-seçimi-feature-selection-teknikleri-5cd8cbab7706>, (Erişim Tarihi: 03/05/2022).
- [24] Web 11, (2020), <https://yigitsener.medium.com/makine-öğrenmesinde-değişken-seçimi-feature-selection-yazı-serisi-gömülü-embedded-yöntemler-c23293915b39>, (Erişim Tarihi: 03/05/2022).
- [25] Web 12, (2022), https://tr.wikipedia.org/wiki/Markov_zinciri, (Erişim Tarihi: 06/06/2022).
- [26] Web 13, (2022), https://en.wikipedia.org/wiki/Bayes%27_theorem, (Erişim Tarihi: 06/06/2022).
- [27] Web 14, (2018), <https://towardsdatascience.com/introduction-to-bayesian-networks-81031eed94e>, (Erişim Tarihi: 06/06/2022).
- [28] Heckerman D., Maxwell C., David M., Christopher M., Rounthwaite R., Kadie C., (2000), "Dependency Networks for Inference, Collaborative Filtering, and Data Visualization", Journal of Machine Learning Research, 49-75.
- [29] Web 15, (2022), [https://en.wikipedia.org/wiki/Dependency_network_\(graphical_model\)](https://en.wikipedia.org/wiki/Dependency_network_(graphical_model)), (Erişim Tarihi: 06/05/2022).
- [30] Demiriz A., (2022), “DiscoVars: A New Data Analysis Perspective - Application in Variable Selection for Clustering”, Teknik Rapor, Endüstri Mühendisliği, Gebze Teknik Üniversitesi.

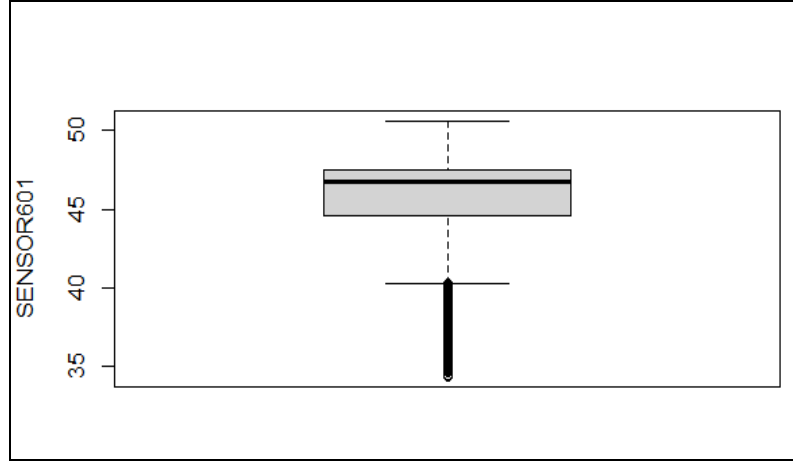
- [31] Web 16, (2007), <https://cloud.google.com/>, (Eriřim Tarihi: 4/06/2022).
- [32] Fox J., Andersen R., (2005), “Using the R Statistical Computing Environment to Teach Social Statistics Courses”, McMaster University.
- [33] Web 17, (2022), [https://en.wikipedia.org/wiki/R_\(programming_language\)](https://en.wikipedia.org/wiki/R_(programming_language)), (Eriřim Tarihi: (25/06/2022)).

ÖZGEÇMİŞ

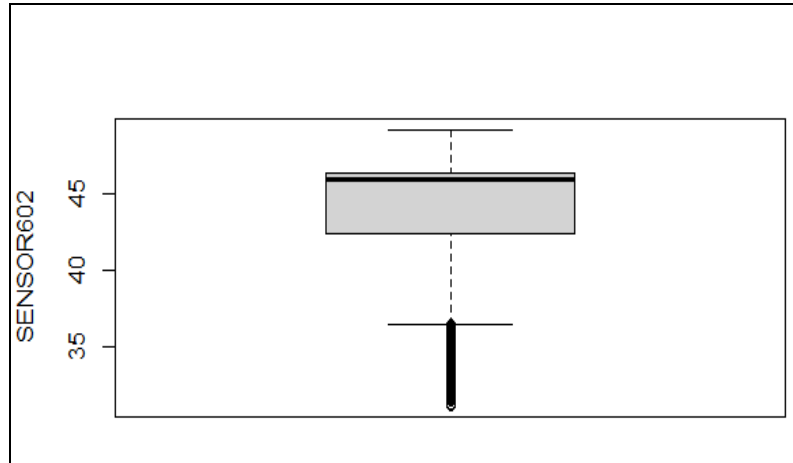
Merve Duman Almanassra, 2013 yılında başlamış olduđu Atatürk Üniversitesi Mühendislik Fakültesi Endüstri Mühendisliđi Bölümünü 2017 yılında tamamlamıştır. 2019 yılında Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliđi Anabilim Dalında yüksek lisans eğitime başlamıştır. 2019 yılından bu yana Şişecam şirketinde İş Zekası ve Veri Analisti Uzman Yardımcısı olarak görev yapmaktadır.

EKLER

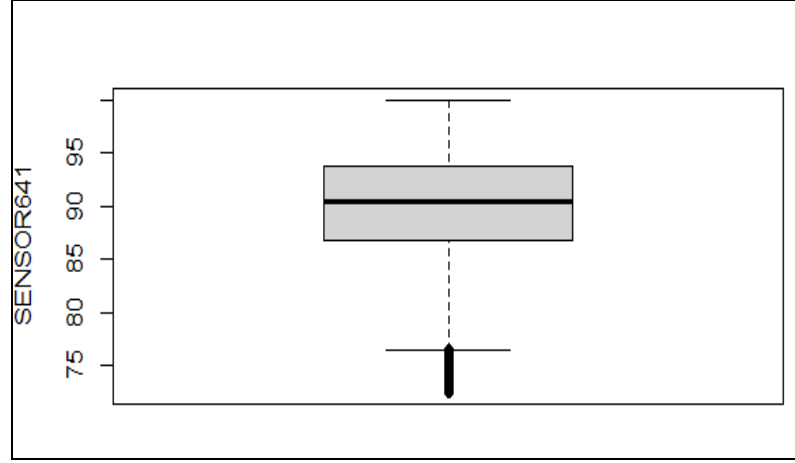
Ek A: Kutu Grafiği



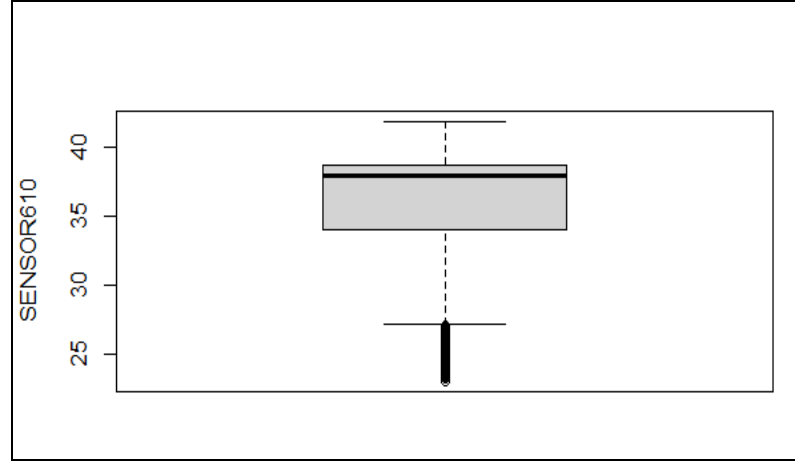
Şekil A1.1: SENSOR601 Kutu grafiği.



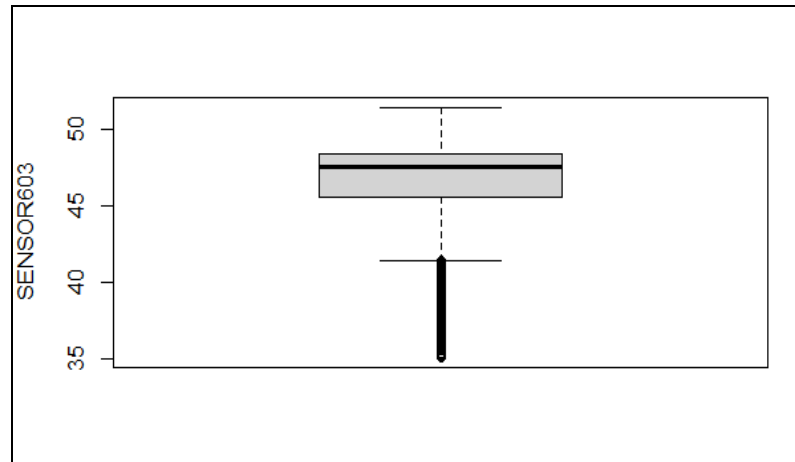
Şekil A1.2: SENSOR602 Kutu grafiği.



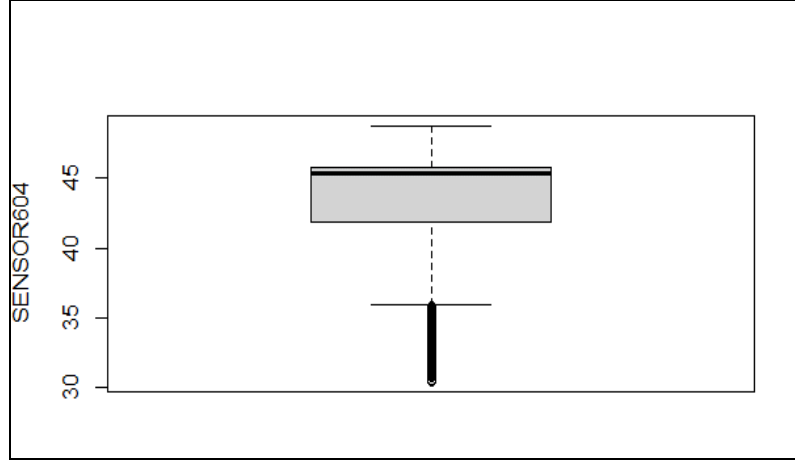
Şekil A1.3: SENSOR641 Kutu grafiği.



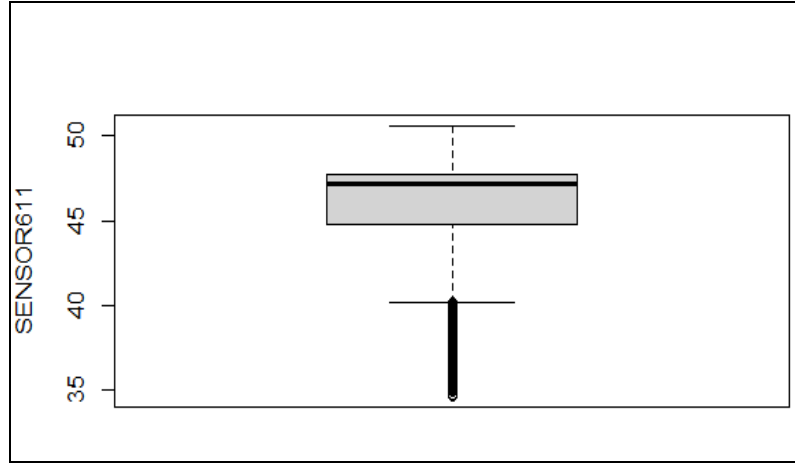
Şekil A1.4: SENSOR610 Kutu grafiği.



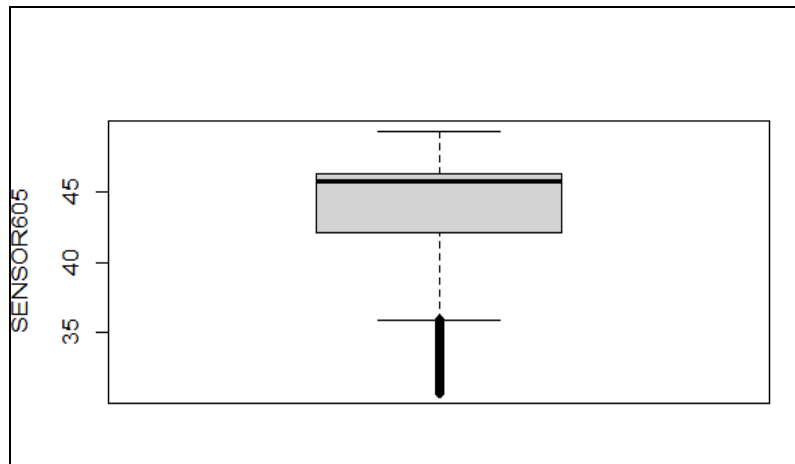
Şekil A1.5: SENSOR603 Kutu grafiği.



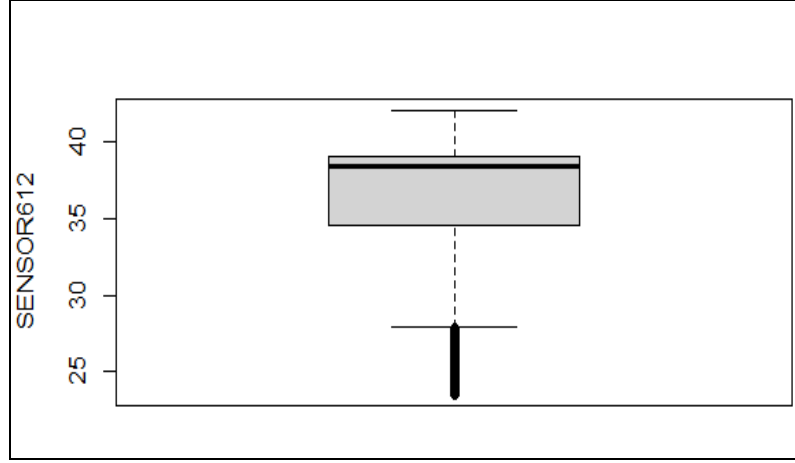
Şekil A1.6: SENSOR604 Kutu grafiği.



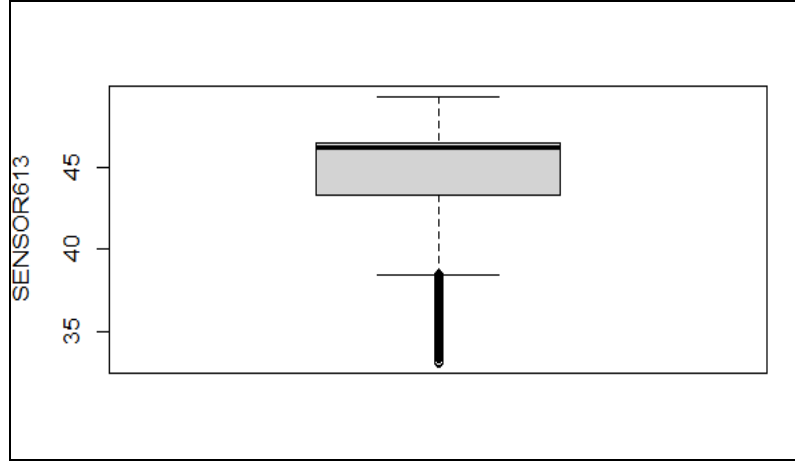
Şekil A1.7: SENSOR611 Kutu grafiği.



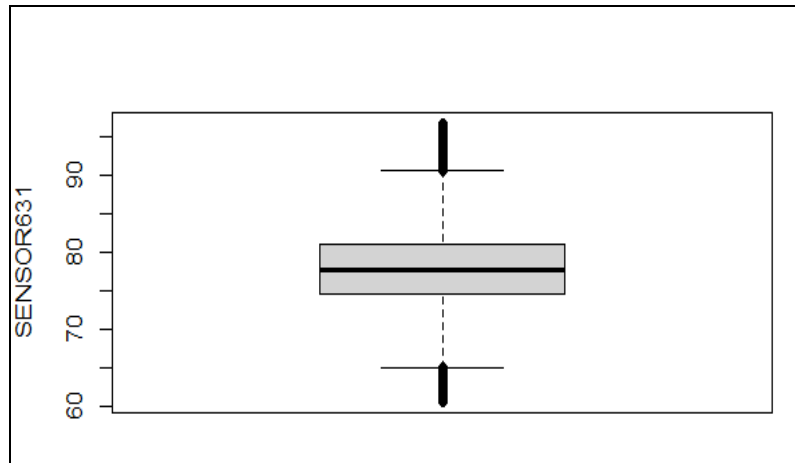
Şekil A1.8: SENSOR605 Kutu grafiği.



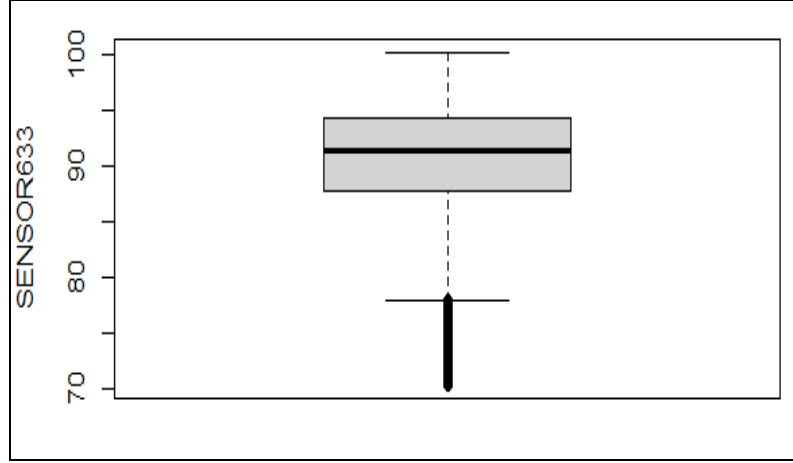
Şekil A1.9: SENSOR612 Kutu grafiği.



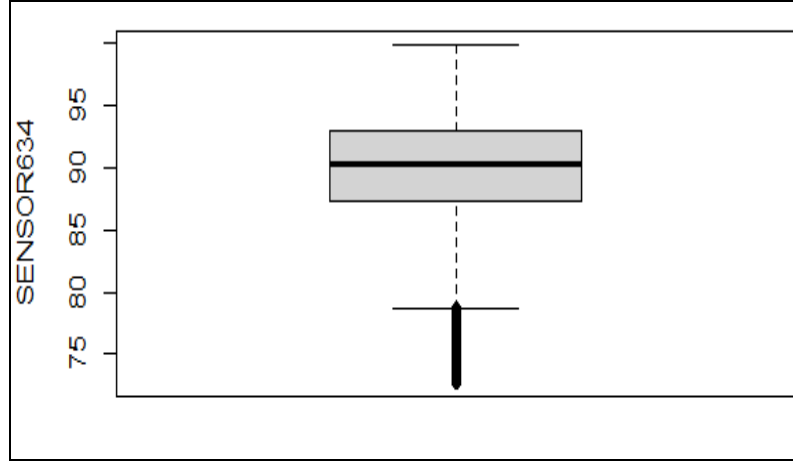
Şekil A1.10: SENSOR613 Kutu grafiği.



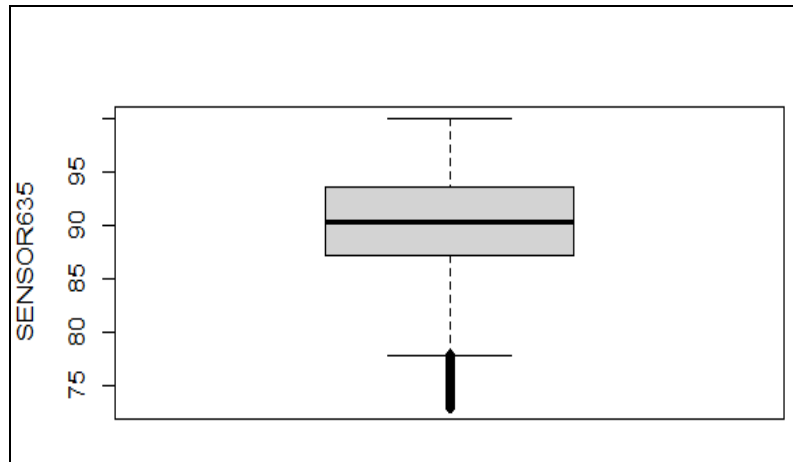
Şekil A1.11: SENSOR631 Kutu grafiği.



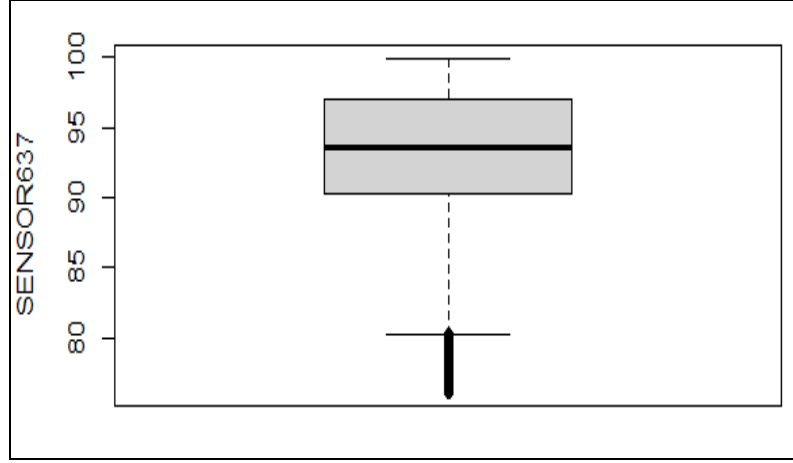
Şekil A1.12: SENSOR633 Kutu grafiği.



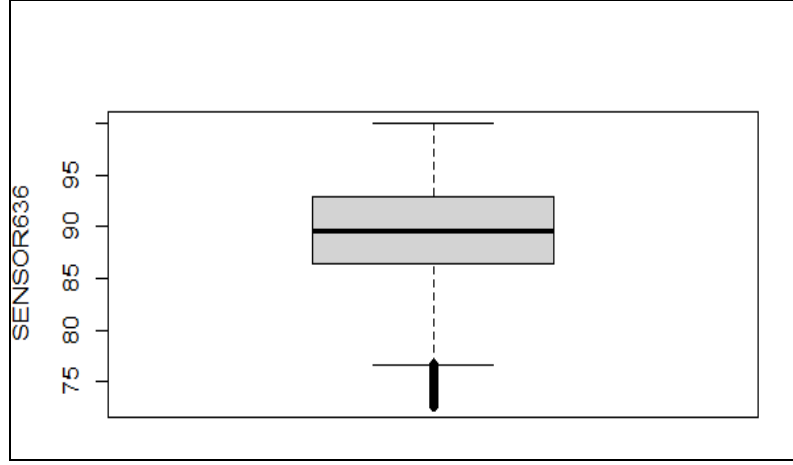
Şekil A1.13: SENSOR634 Kutu grafiği.



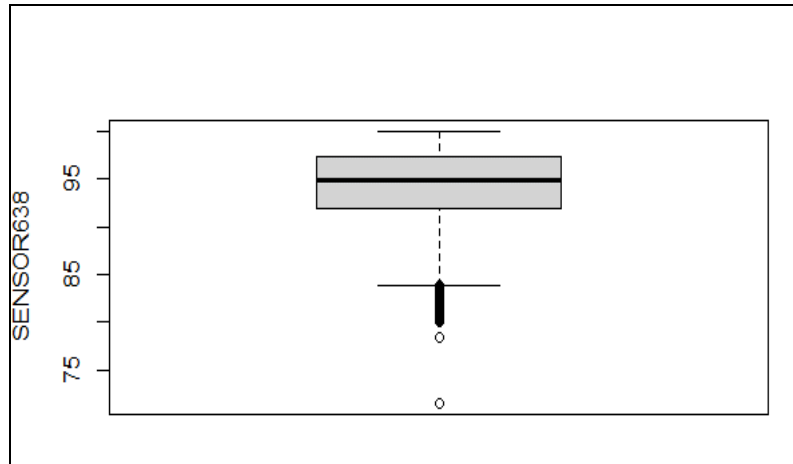
Şekil A1.14: SENSOR635 Kutu grafiği.



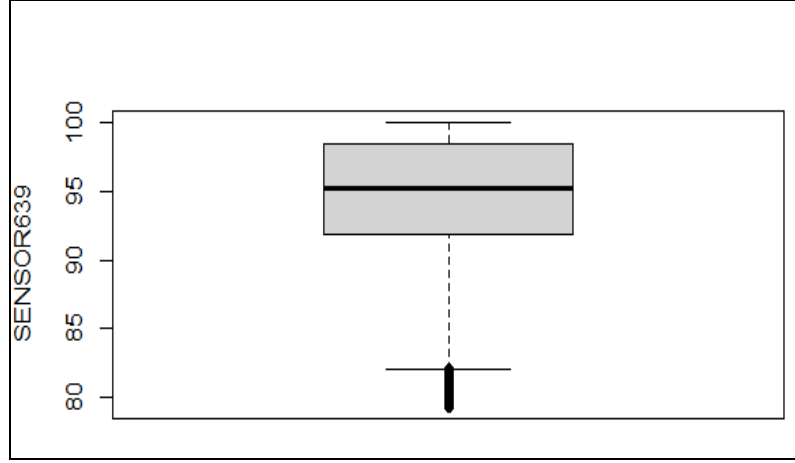
Şekil A1.15: SENSOR637 Kutu grafiği.



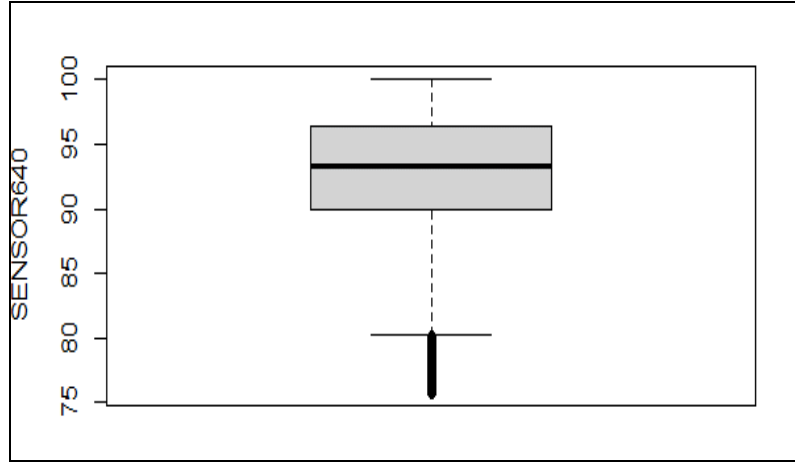
Şekil A1.16: SENSOR636 Kutu grafiği.



Şekil A1.17: SENSOR638 Kutu grafiği.

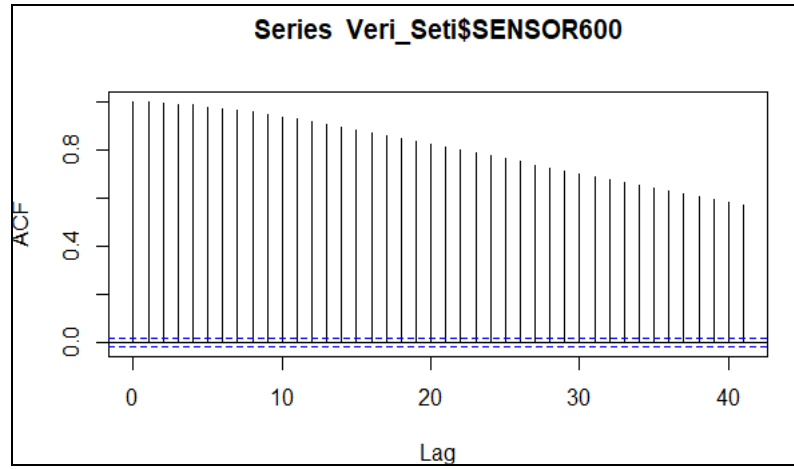


Şekil A1.18: SENSOR639 Kutu grafiği.

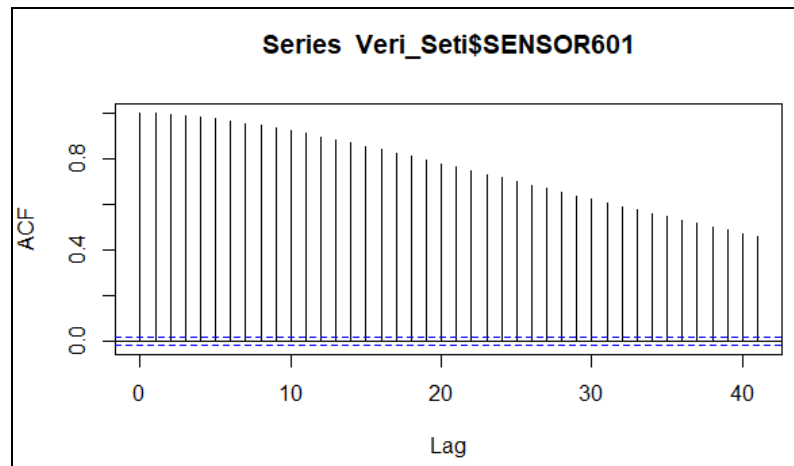


Şekil A1.19: SENSOR640 Kutu grafiği.

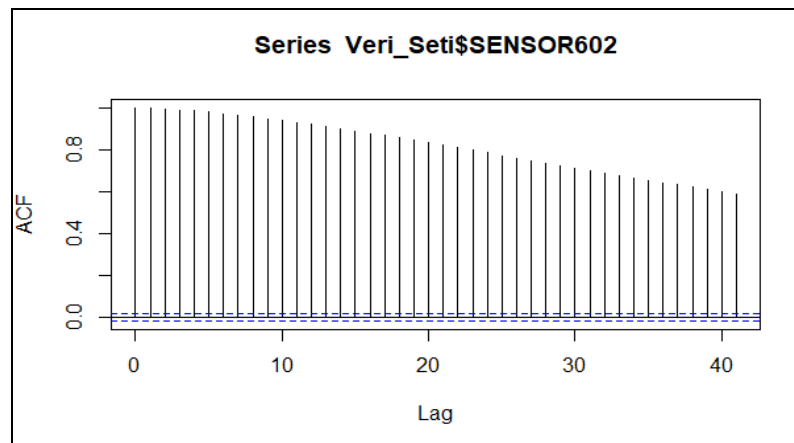
Ek B: Otokorelasyon Grafiđi



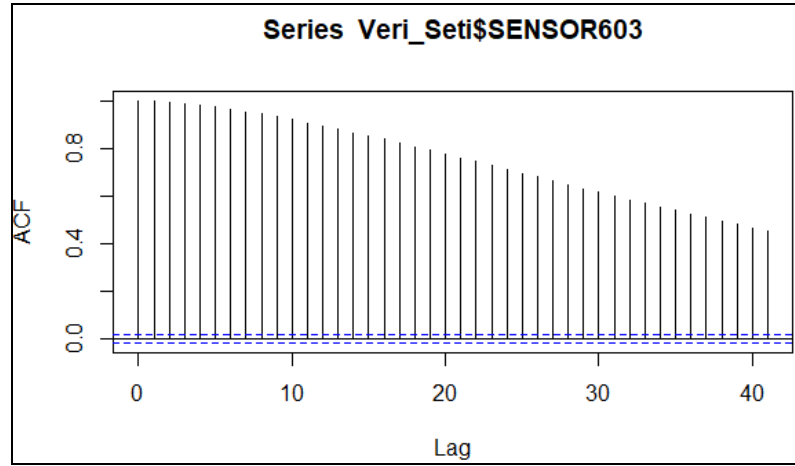
Şekil B1.1: SENSOR600 Otokorelasyon grafiđi.



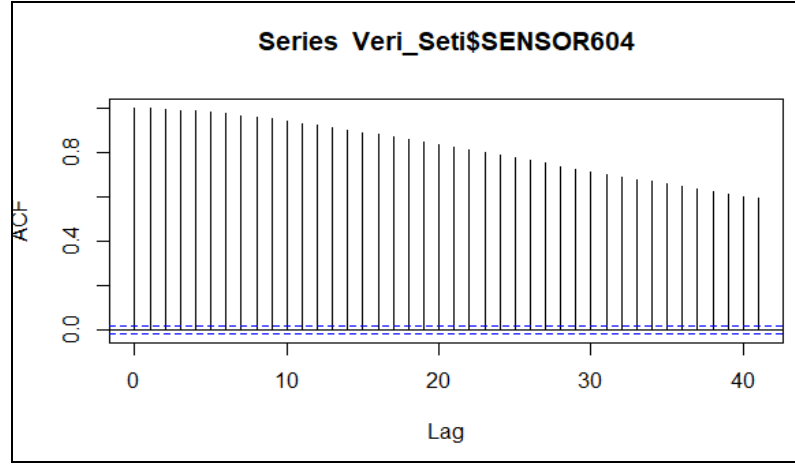
Şekil B1.2: SENSOR601 Otokorelasyon grafiđi.



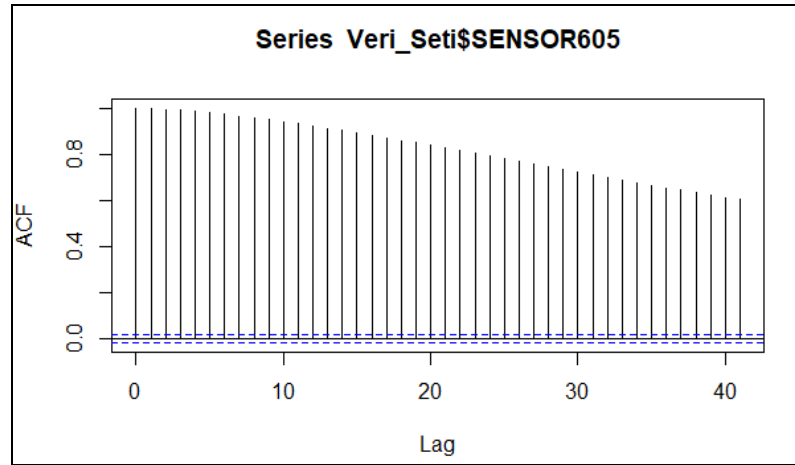
Şekil B1.3: SENSOR602 Otokorelasyon grafiđi.



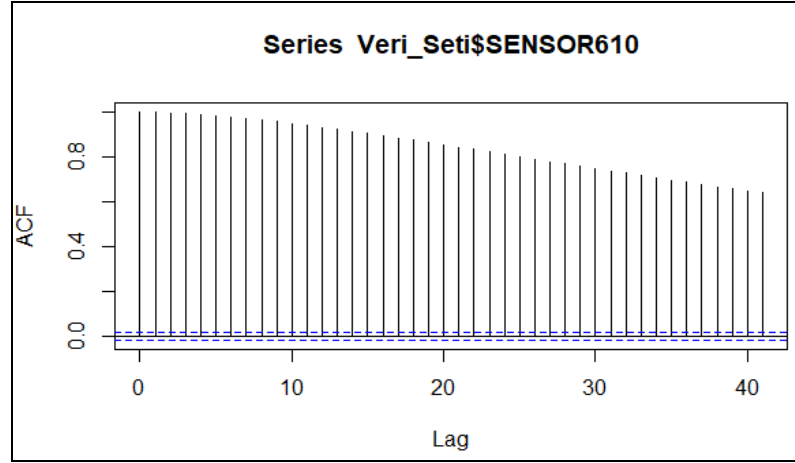
Şekil B1.4: SENSOR603 Otokorelasyon grafiği.



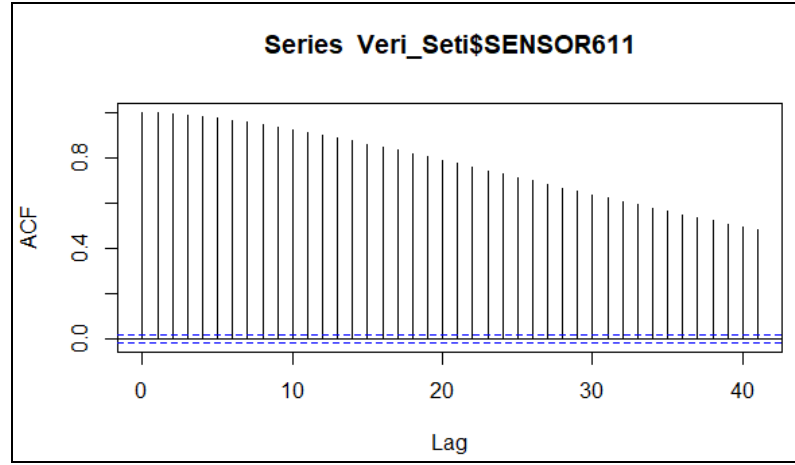
Şekil B1.5: SENSOR604 Otokorelasyon grafiği.



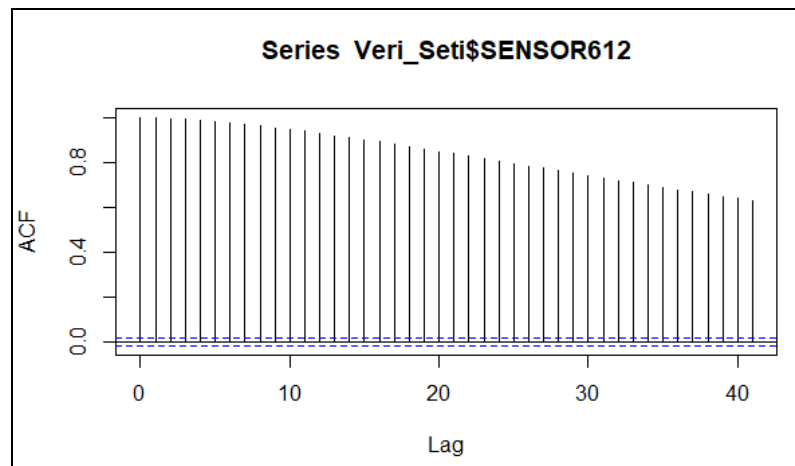
Şekil B1.6: SENSOR605 Otokorelasyon grafiği.



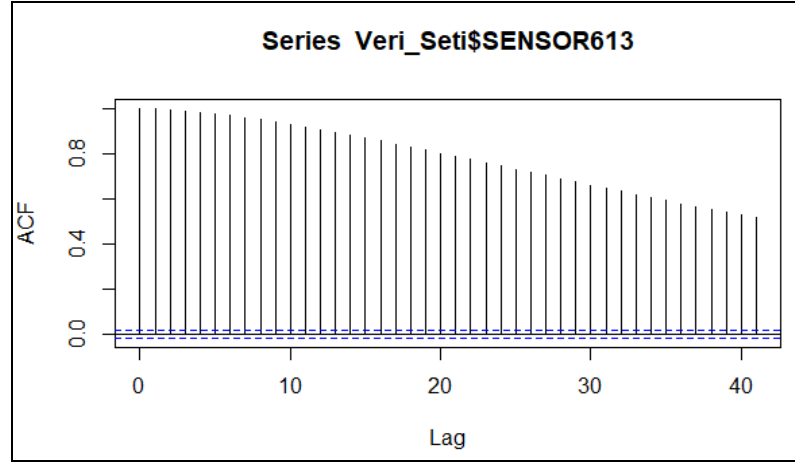
Şekil B1.7: SENSOR610 Otokorelasyon grafiği.



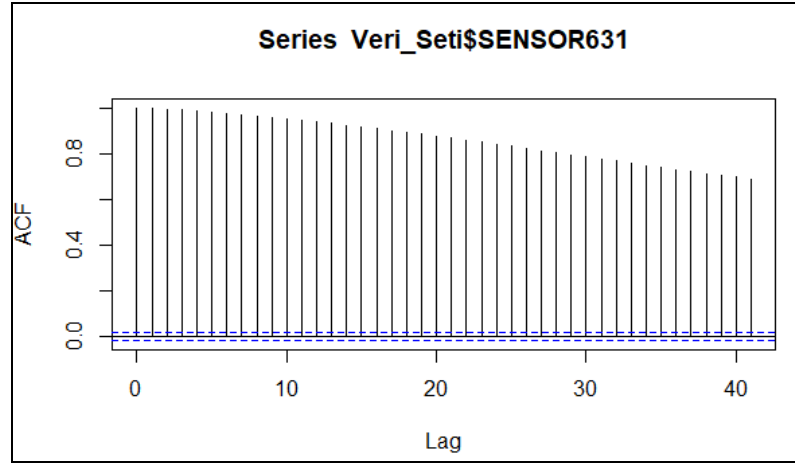
Şekil B1.8: SENSOR611 Otokorelasyon grafiği.



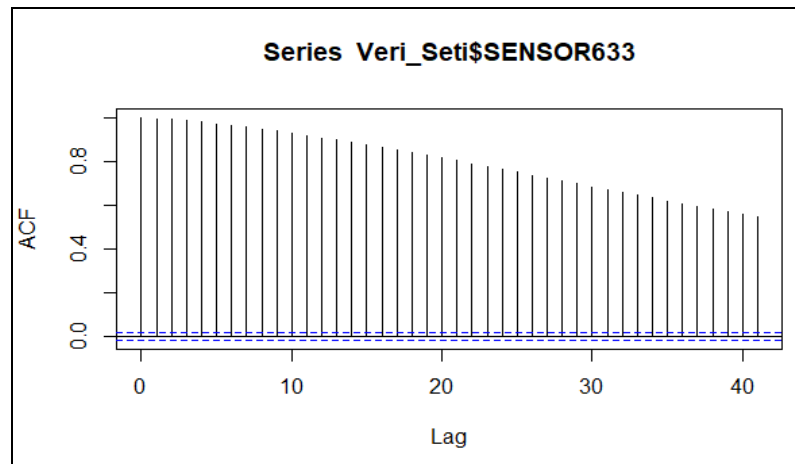
Şekil B1.9: SENSOR612 Otokorelasyon grafiği.



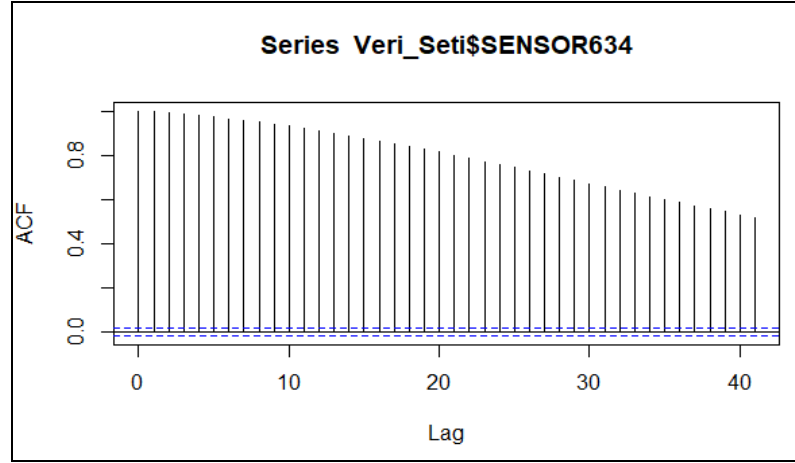
Şekil B1.10: SENSOR613 Otokorelasyon grafiği.



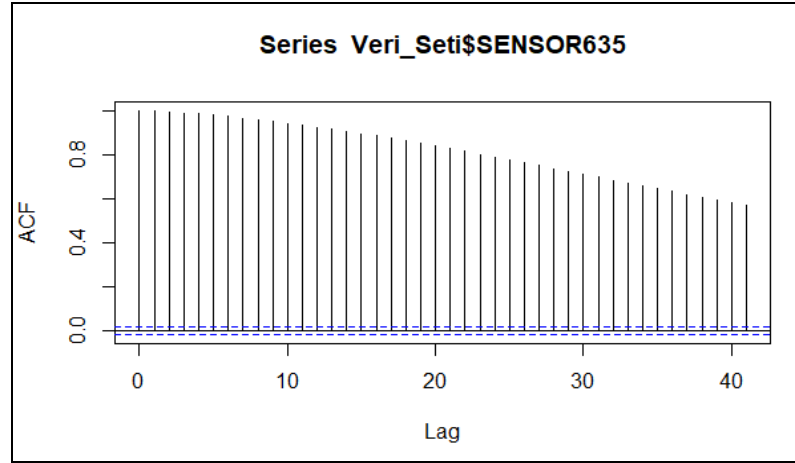
Şekil B1.11: SENSOR631 Otokorelasyon grafiği.



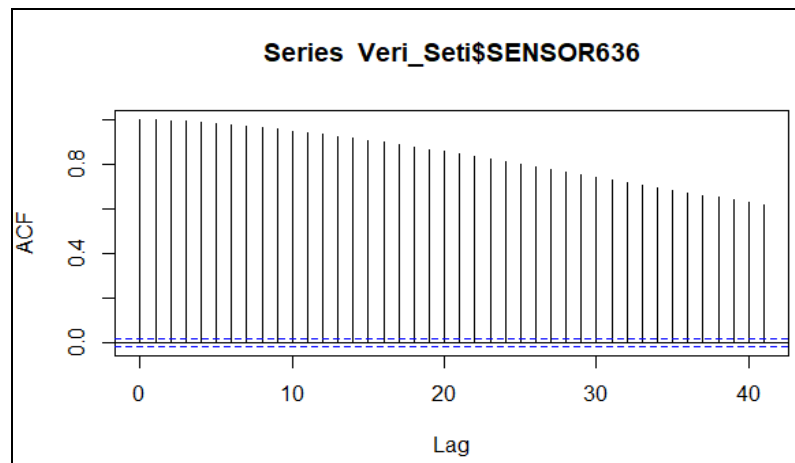
Şekil B1.12: SENSOR633 Otokorelasyon grafiği.



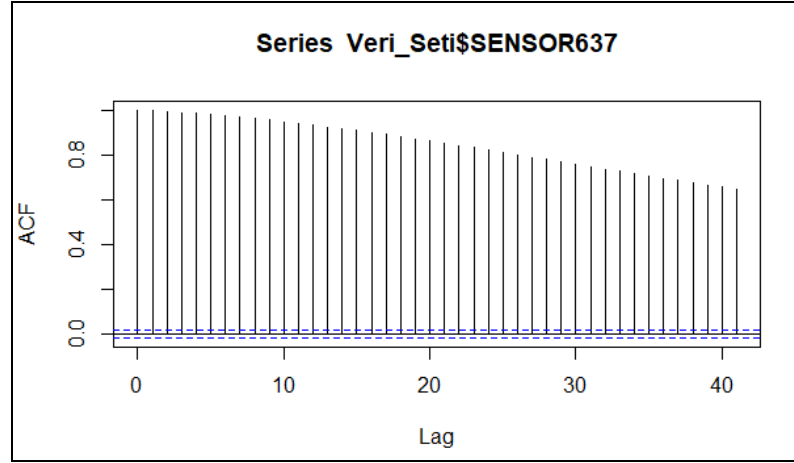
Şekil B1.13: SENSOR634 Otokorelasyon grafiği.



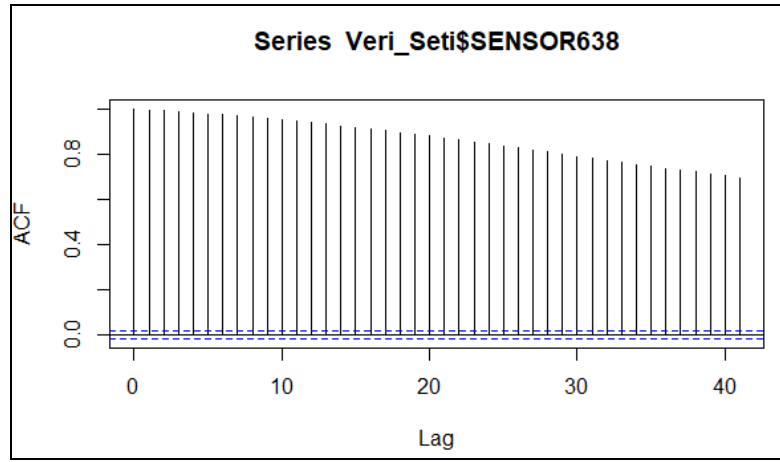
Şekil B1.14: SENSOR635 Otokorelasyon grafiği.



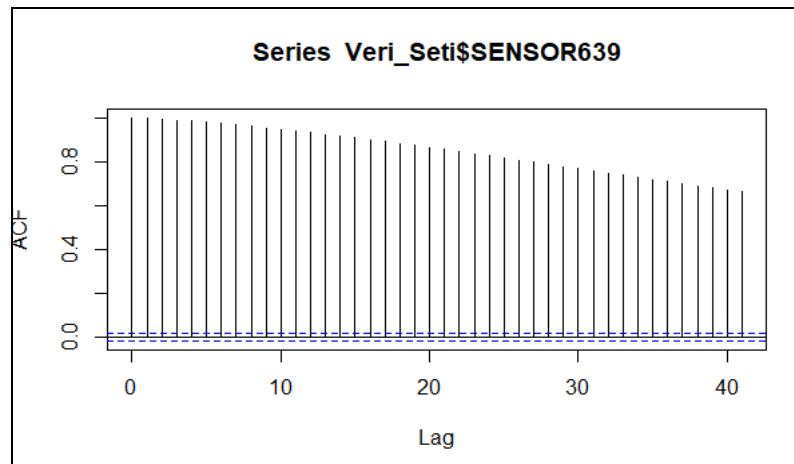
Şekil B1.15: SENSOR636 Otokorelasyon grafiği.



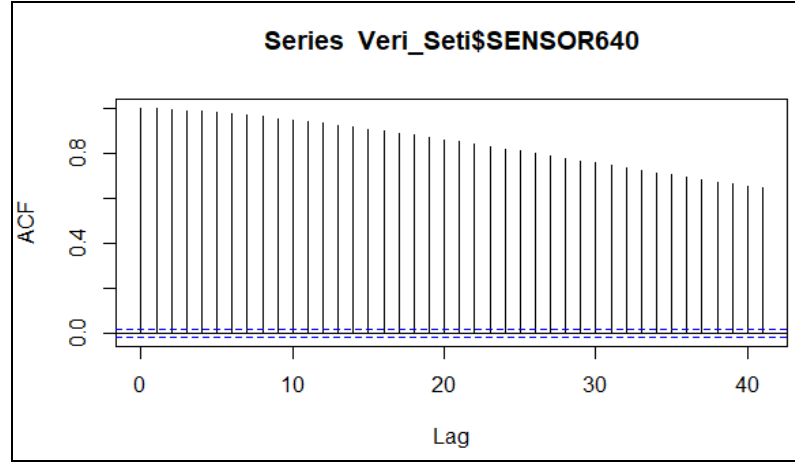
Şekil B1.16: SENSOR637 Otokorelasyon grafiği.



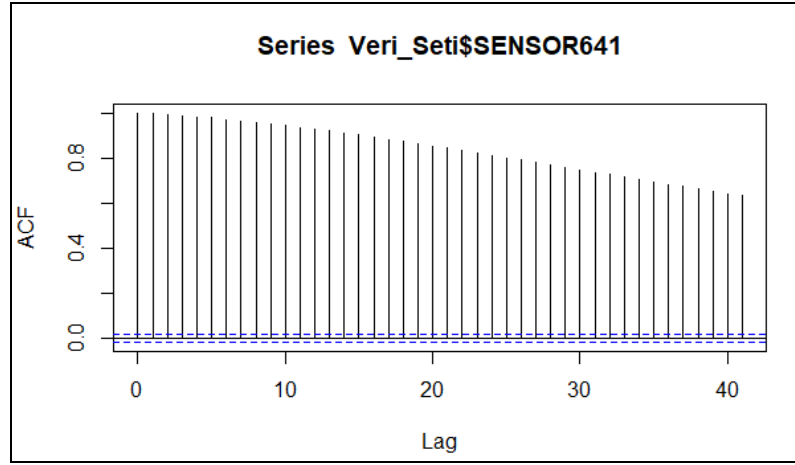
Şekil B1.17: SENSOR638 Otokorelasyon grafiği.



Şekil B1.18: SENSOR639 Otokorelasyon grafiği.

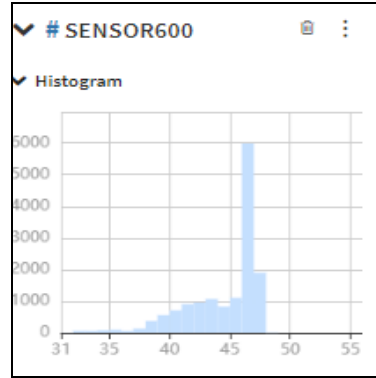


Şekil B1.19: SENSOR640 Otokorelasyon grafiği.

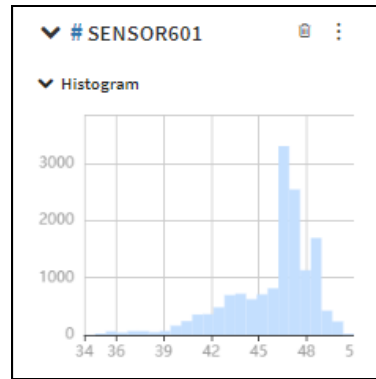


Şekil B1.20: SENSOR641 Otokorelasyon grafiği.

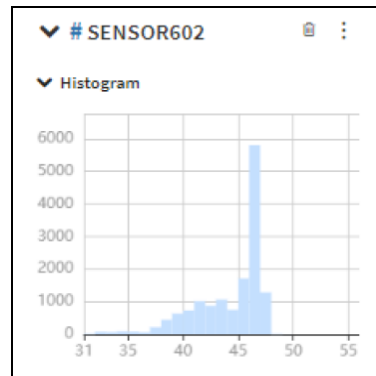
Ek C: Histogram Grafiđi



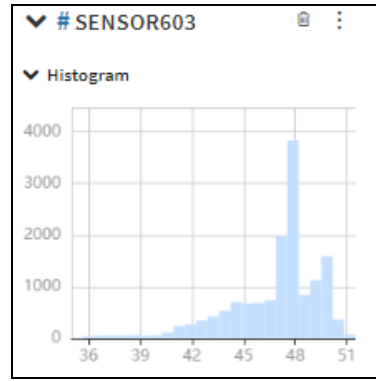
Şekil C1.1: SENSOR600 Histogram grafiđi.



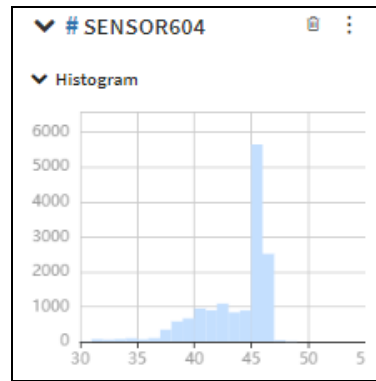
Şekil C1.2: SENSOR601 Histogram grafiđi.

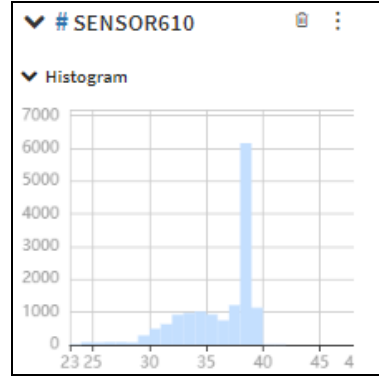


Şekil C1.3: SENSOR602 Histogram grafiđi.

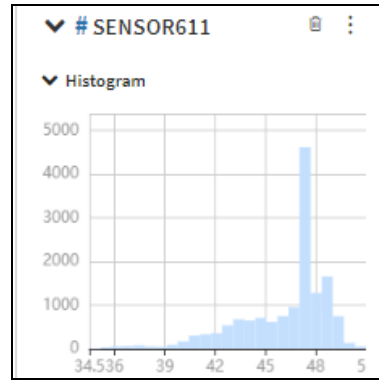


Şekil C1.4: SENSOR603 Histogram grafiği.

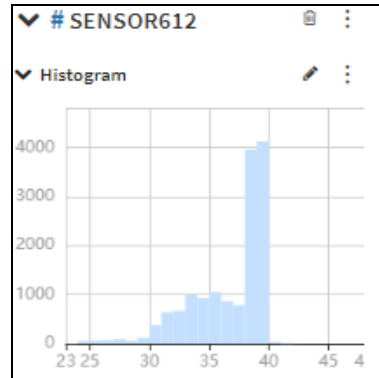




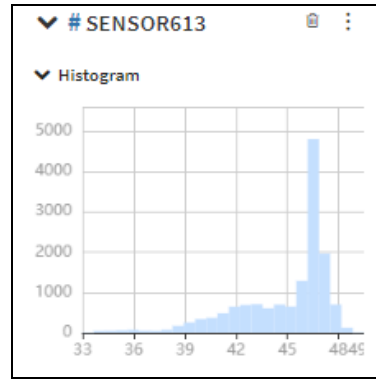
Şekil C1.7: SENSOR610 Histogram grafiği.



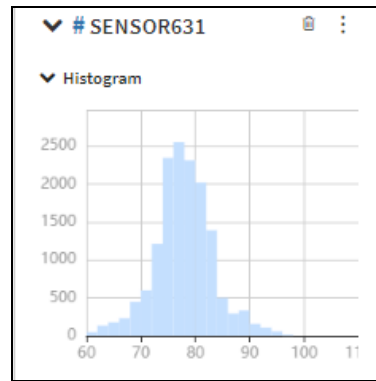
Şekil C1.8: SENSOR611 Histogram grafiği.



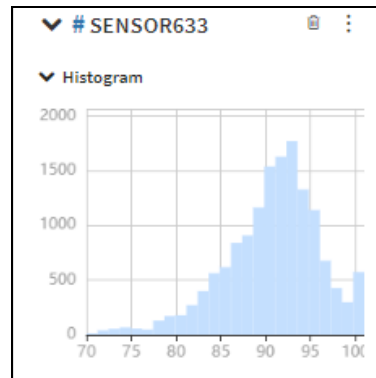
Şekil C1.9: SENSOR612 Histogram grafiği.



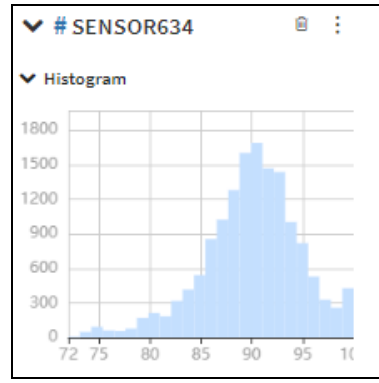
Şekil C1.10: SENSOR613 Histogram grafiği.



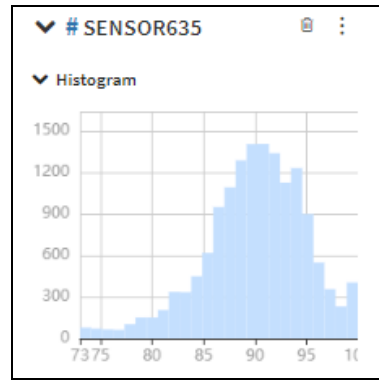
Şekil C1.11: SENSOR631 Histogram grafiği.



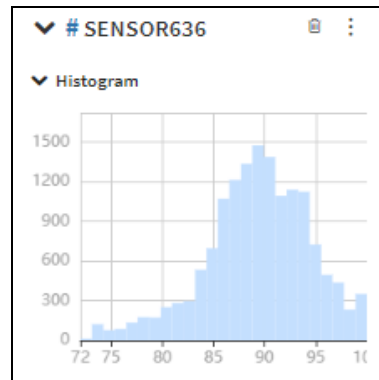
Şekil C1.12: SENSOR633 Histogram grafiği.



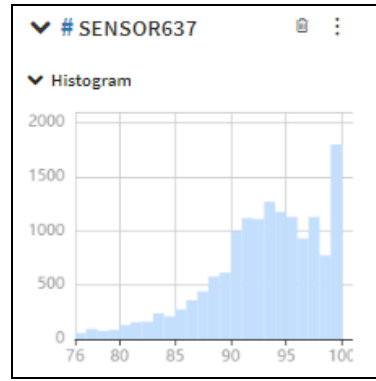
Şekil C1.13: SENSOR634 Histogram grafiği.



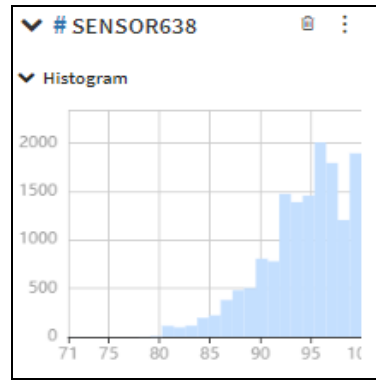
Şekil C1.14: SENSOR635 Histogram grafiği.



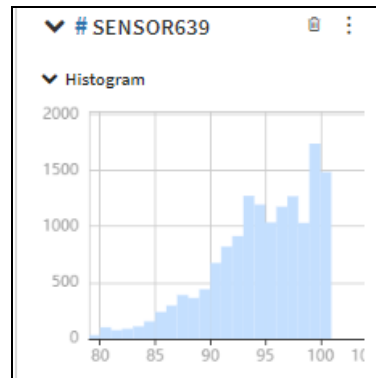
Şekil C1.15: SENSOR636 Histogram grafiği.



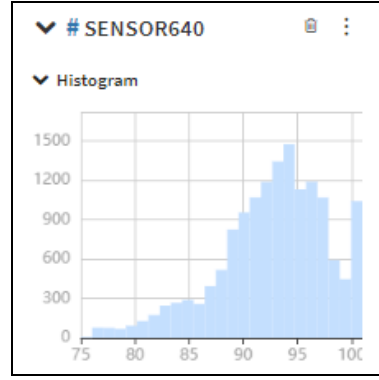
Şekil C1.16: SENSOR637 Histogram grafiği.



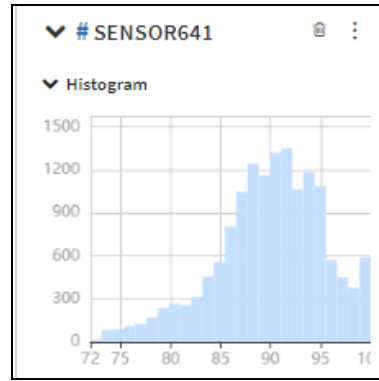
Şekil C1.17: SENSOR638 Histogram grafiği.



Şekil C1.18: SENSOR639 Histogram grafiği.



Şekil C1.19: SENSOR640 Histogram grafiği.



Şekil C1.20: SENSOR641 Histogram grafiği.