**2017-2018 BAHAR DÖNEMİ**

**İST 347 İSTATİSTİKSEL ANALİZ UYGULAMALARI FİNAL SINAVI**

**Giriş:** Çözülen probleme veri çözümleme açısından bakarsak R programının kullanımını daha iyi anlamış ve çözümleme yapma, bunları yorumlama pratiğimizi geliştirmiş olacağız. Veri içeriği yönünden bakarsak verilerimiz hava kaliteleri ile ilgili bilgileri içermektedir ve günümüzde hava kirliliği ve bunların sonuçları günden güne artmaktadır. Bu analizleri yaparak bunun doğru ya da yanlış olduğunu, doğruysa ne seviyeye kadar geldiğini anlamış olacağız.

**Gerekli Paketler:**

**Rcmdr Paketi:** Rcmdr paketi açılımı R Commander olan güçlü bir pakettir. Basitçe anlatmak gerekirse istatistiksel her türlü analizi tek tuşla yapabildiğimiz ve aynı zamanda yazılan kodları bize vererek kendimizi geliştirmemize olanak veren bir pakettir. İleride ayrıntı verilecektir.

**Veri Hazırlama:** Verilerimiz bize hava kalitesi hakkında bilgi vermektedir. Yakın zamanda hava içerisinde bulunan ozon, solar, rüzgar hızı, sıcaklık (Fahrenheit cinsinden) ve bu ölçümlerin hangi aylarda yapıldığı gibi bilgiler vardır. Aynı zamanda her bir veri kümesinin her bir değerinin ortalamadan fazla veya az olduğuyla ilgili 4 farklı değişken daha oluşturulmuştur.

Veri yüklemesi yaparken R-Studio programına tipik text dosyası kullanarak veri yüklemesi gerçekleştirilmiştir. Düzenleme yaparken attach fonksiyonu kullanılarak her bir değişken üzerinden işlem yapılması sağlanmıştır. Aynı zamanda ‘ay’ değişkeni üzerinde ki sayısal verileri kategorik veri haline getirmek veri analizi için önemliydi. Bunu yaparken algoritmayı basitçe anlatmak gerekirse değişkende ki her bir sayının ‘ay’ karşılığı if-else basamak yapısı ile yeni bir diziye aktarıldı ve bu dizi verinin üzerine yazdırıldı.

**Ozon Değişkeni:** Bu değişken hava içerisinde bulunan ozon miktarı hakkında bize bilgi vermektedir. Sayısal verilerden oluşmaktadır dolayısıyla üzerinde tanımlayıcı istatistik hesaplaması yapabiliriz. Bundan sonra ki bütün değişkenler içinde aynı tanımlayıcı istatistikler koyulacaktır.

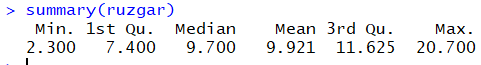


Görüldüğü üzere en basit haliyle ozon miktarına ilişkin ortalama 42.29 olarak bulunmuştur.

**Solar Değişkeni:** Bu değişken bize havadaki solar radrasyon değerini vermektedir. Sayısal verilerden oluşmaktadır.



**Rüzgar Değişkeni:** Bu değişken tahminimce kilometre cinsinden bize rüzgar hızını vermektedir. Sayısal verilerden oluşmaktadır.



**Sıcaklık Değişkeni:** Bu değişken bize Fahrenheit cinsinden sıcaklık değerlerini vermektedir. Sayısal verilerden oluşmaktadır.



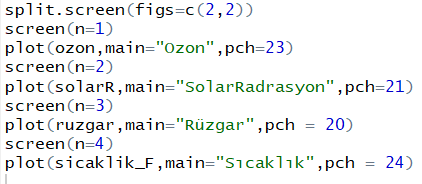
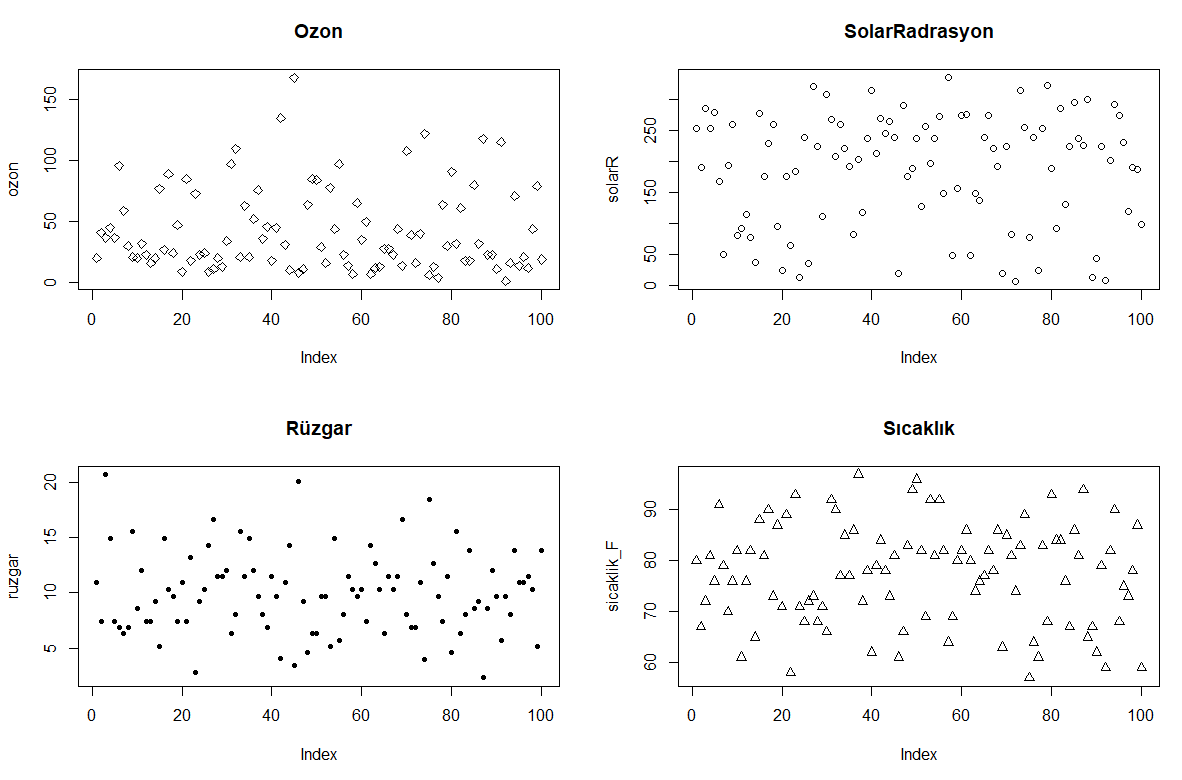
**Ay Değişkeni:** Bu değişken bize başlangıçta sayısal veriler şeklinde verilmiştir. Fakat sonradan kategorik veri haline getirilmiş ve verilerin hangi aylar hakkında bilgi verdiğini gösteren bir değişken haline gelmiştir.

**‘Yüksek’ Değişkenleri:** Her bir yüksek değişkeni kendi değişkeninin ortanca değerinin üstündeyse 1 altındaysa 0 değerini verecek şekilde oluşturulmuştur. Burada 1 ve 0, true ve false değerler olarak kabul edilebilir.

**Analiz Kısmı:**

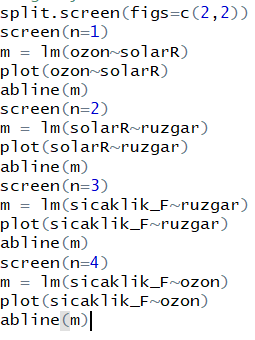
İlk olarak elimizde bulunan ve kategorik olmayan 4 tane değişkenin herbirinin ayrı ayrı dağılım grafiklerine bakalım. Bunu yaparken 4 grafiği de aynı pencereye yerleştireceğim ve bunu diğer analizler içinde yapacağım.

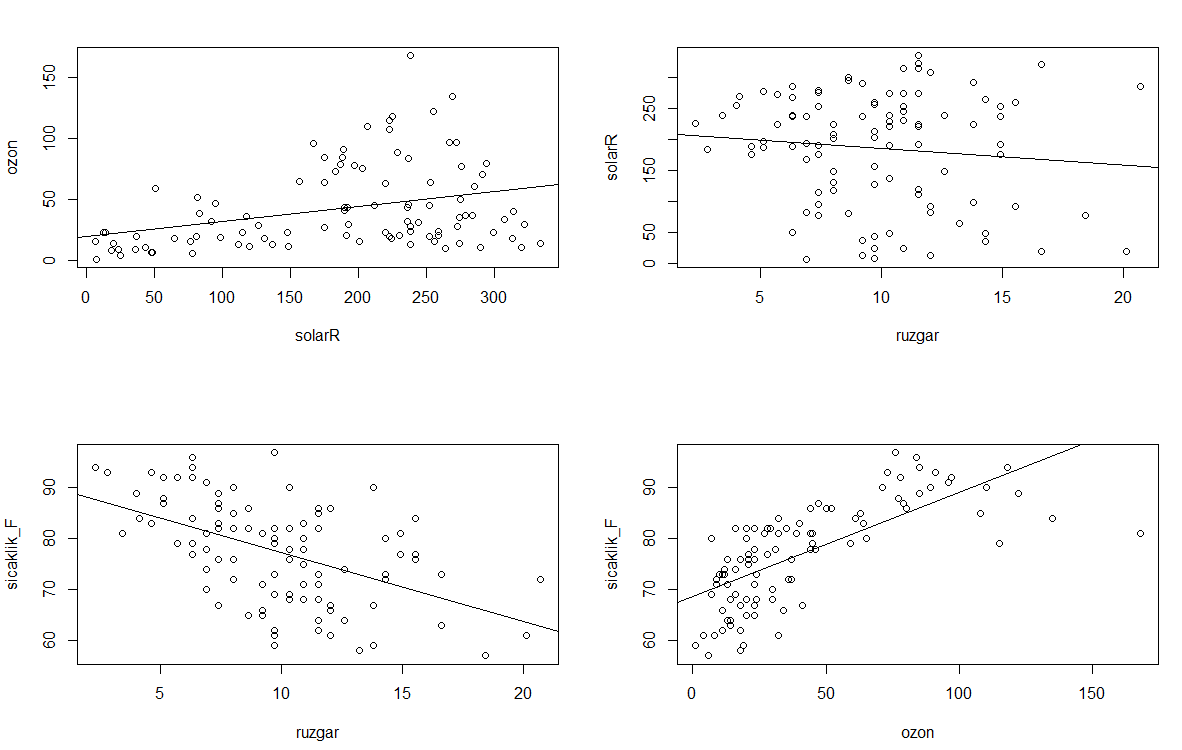
Grafik diğer sayfada.

****

Görüldüğü üzere ozon, solar radrasyon, rüzgar ve sıcaklık değişkenlerinin dağılım grafiği karşımıza gelmiştir. Dağılım grafikleri ile ilgili yapılacak çok fazla bir yorum bulunmamaktadır. Bu grafikler bize verileri görselleştirerek anlaşılmasında kolaylık sağlamaktadır.

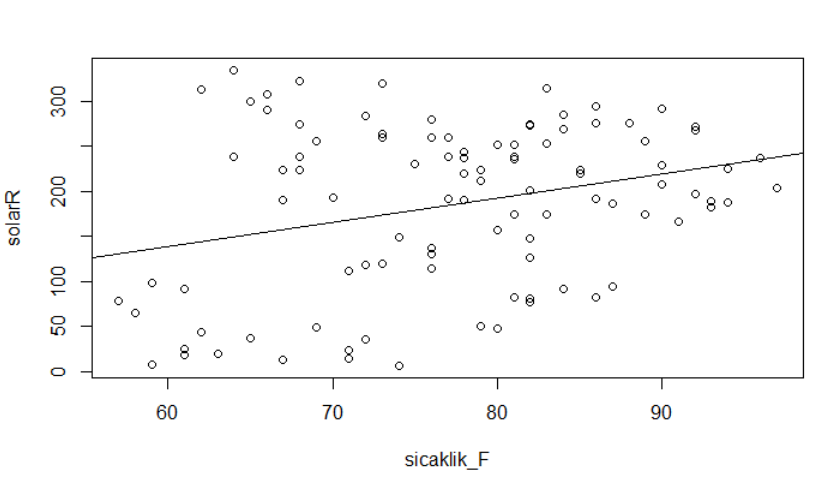
Şimdi birkaç tane değişken arasında regresyon çizgilerini inceleyip bağlantı olup olmadığına bakalım.





Yukarıda 4 farklı ikili üzerinden regresyon çizgileri çekilmiştir. En belirgin farkların sıcaklık-rüzgar ve sıcaklık-ozon arasında olduğu söylenebilir. Sıcaklık ve rüzgar arasında negatif bir ilişki söz konusudur. Yani rüzgar arttıkça sıcaklık düşmektedir. Aynı şekilde ozon ve sıcaklık arasında da pozitif bir ilişki söz konusudur. Yani ozon oranı arttıkça sıcaklık artmaktadır. Zaten hava kirliliğinin en belirgin sebebi ozon artışıdır. Bu da sıcaklıkların gittikçe artmasına ve küresel ısınmaya sebep olmaktadır. Burada her iki karşılaştırmada da sıcaklık faktörü ele alınmıştır öyleyse ozon ve rüzgar arasında da negatif bir ilişki olduğu söylenebilir.

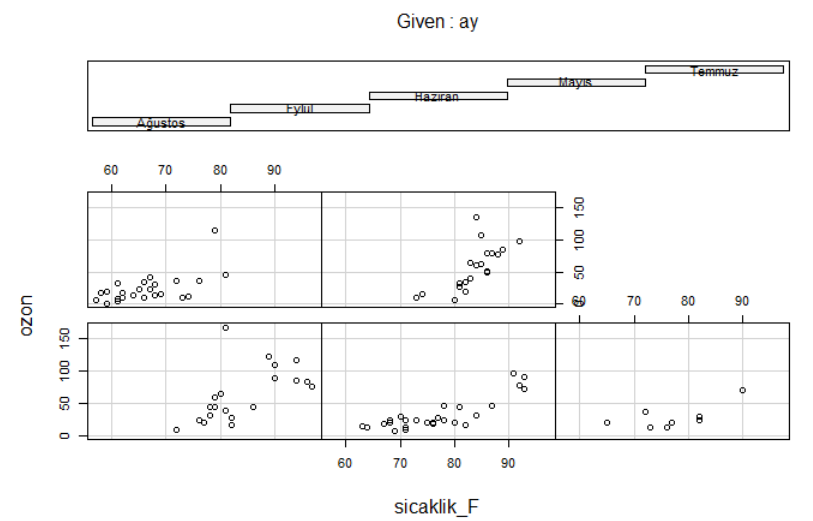
Solar Radyasyonun ilk grafikte görüldüğü üzere ozon ile çok fazla olmasa da pozitif bir ilişkisi olduğu görülmektedir. Buda aynı zamanda sıcaklık ile de aynı derecede çok fazla olmayan pozitif bir ilişkisi olduğu anlamına gelir. Bunu grafik ile güçlendirelim.



Tıpkı ilk grafikte olduğu gibi (n=1) aynı derecede eğime sahip bir regresyon çizgisi görmekteyiz. Bu tarz yorumları yukarıda ki 4 grafik sayesinde yapabiliriz. Buda her bir ikili değişken arasında regresyon çizgisini çekip grafikleri yorumlamamızın gereksiz olduğunu göstermektedir.

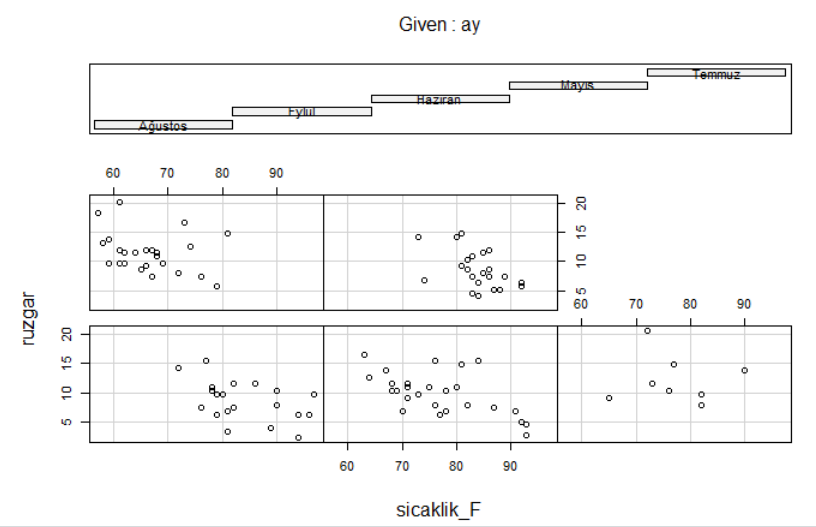
Sıcaklık-Ozon ve Sıcaklık-Rüzgar arasında güçlü ilişki olduğuna göre bunların aylara göre saçılım grafiklerini çizelim.





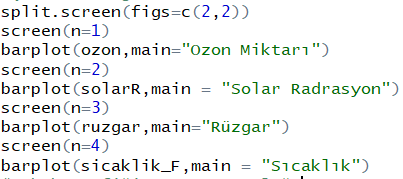
Yukarıda ki grafikte her bir aya göre sıcaklık ve ozon arasında ki saçılım grafiğini görmekteyiz. Burada grafikler önce yukarı ve soldan sağa doğru gitmektedir. İlk sütunu ele alırsak bize haziran ve ağustos aylarında ki grafikleri vermektedir ve haziran ayında sıcaklıkların daha fazla olması ağustos ayına gelindiğinde ozon oranıyla birlikte sıcaklıklarında düştüğü görülmektedir. Aynı grafiği sıcaklık-rüzgar içinde yapalım.

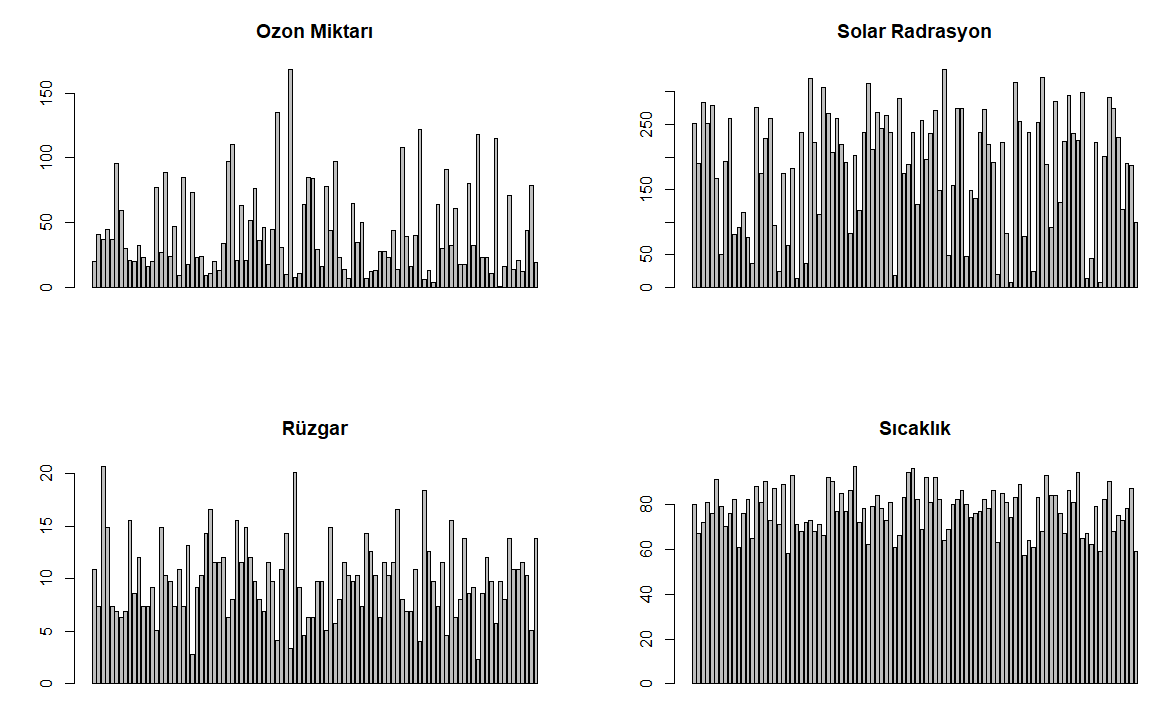




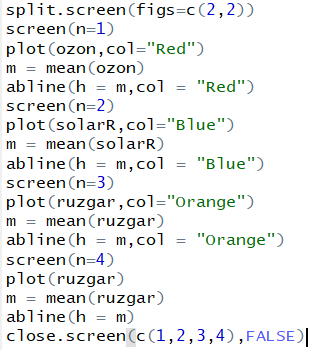
Burada aynı sütunu yorumlarsak sıcaklık haziran ayından ağustos ayına geldiğinde düşmüştür ve bununla beraber ağustos ayında rüzgar hızında artış görülmüştür.

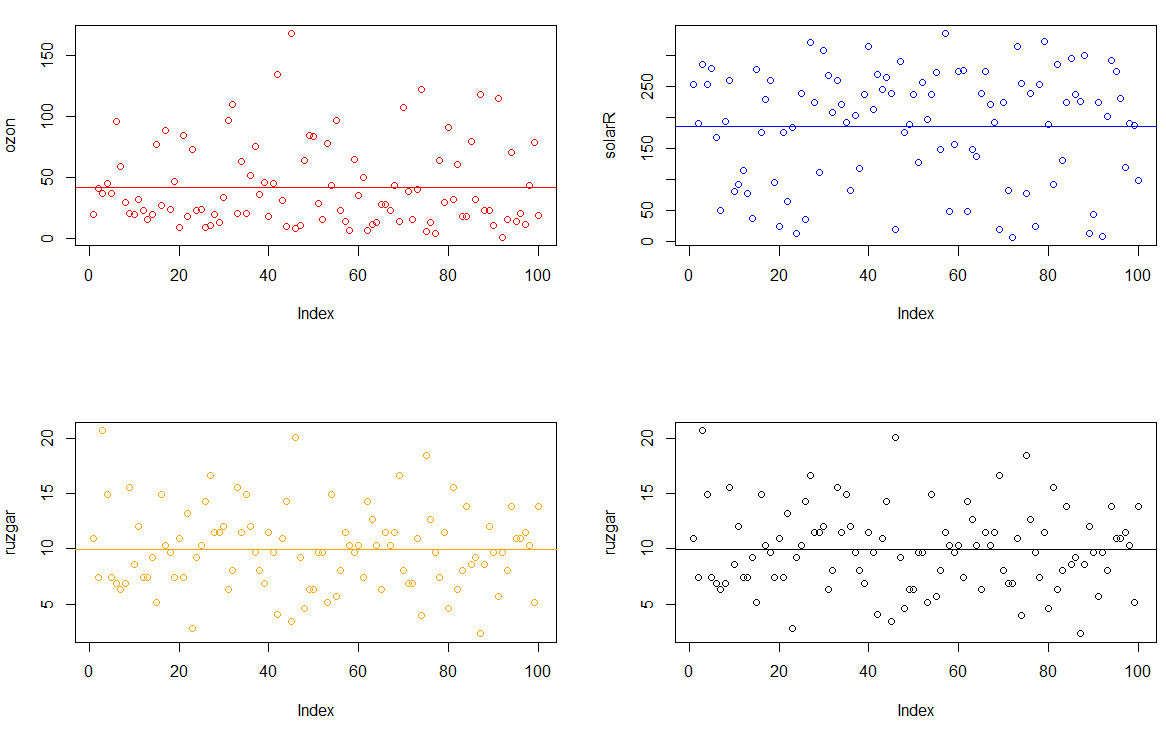
Şimdi elimize bulunan 4 ana değişkenin çubuk grafiklerini çizerek yorumlamayı kolaylaştırabilir hale getirelim.





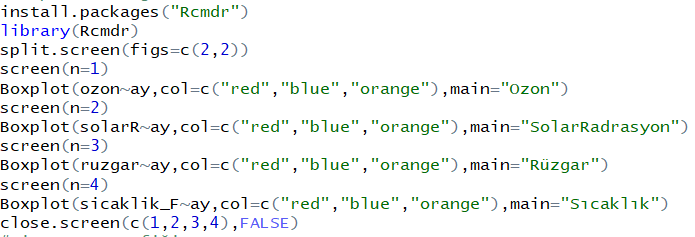
Elimizde ki verilerin ortalamalarının dağılım grafiğinde nerede olduğunu bilmek isteyebiliriz. Bunu elimizde ki 4 ana değişken için uygulayalım.

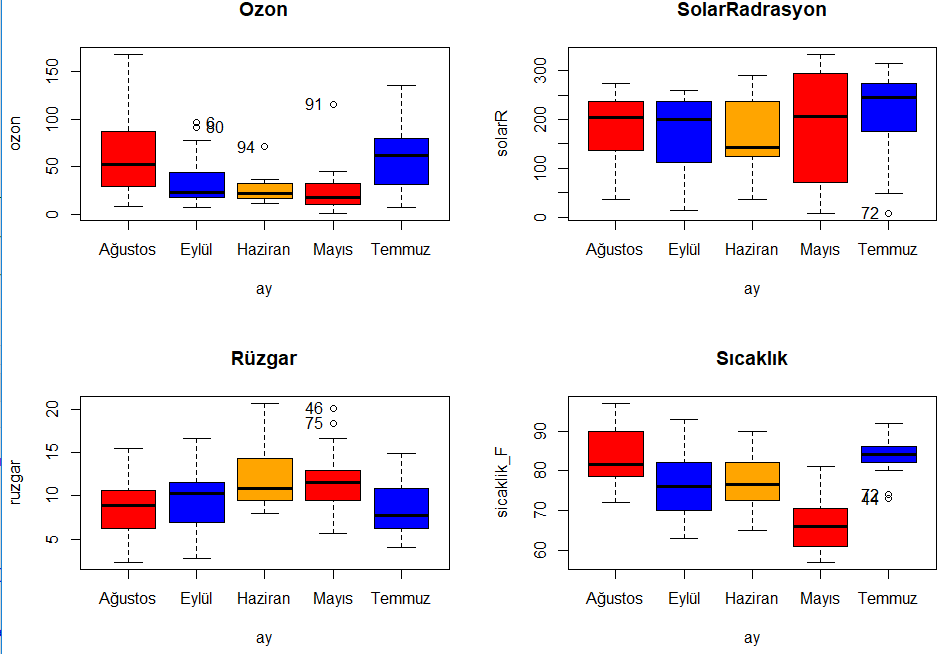




Burada ki çizgiler tam olarak ortalama üzerinden geçmektedir. İlk başta koyduğumuz tanımlayıcı istatistiklerde ki ortalama değerleriyle birebir uyuşmaktadır.

Elimizdeki 4 ana verinin sonradan kategorik veri haline getirdiğimiz ay verisine göre boxplot grafiklerini çizelim.

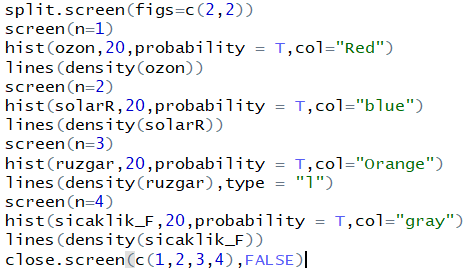


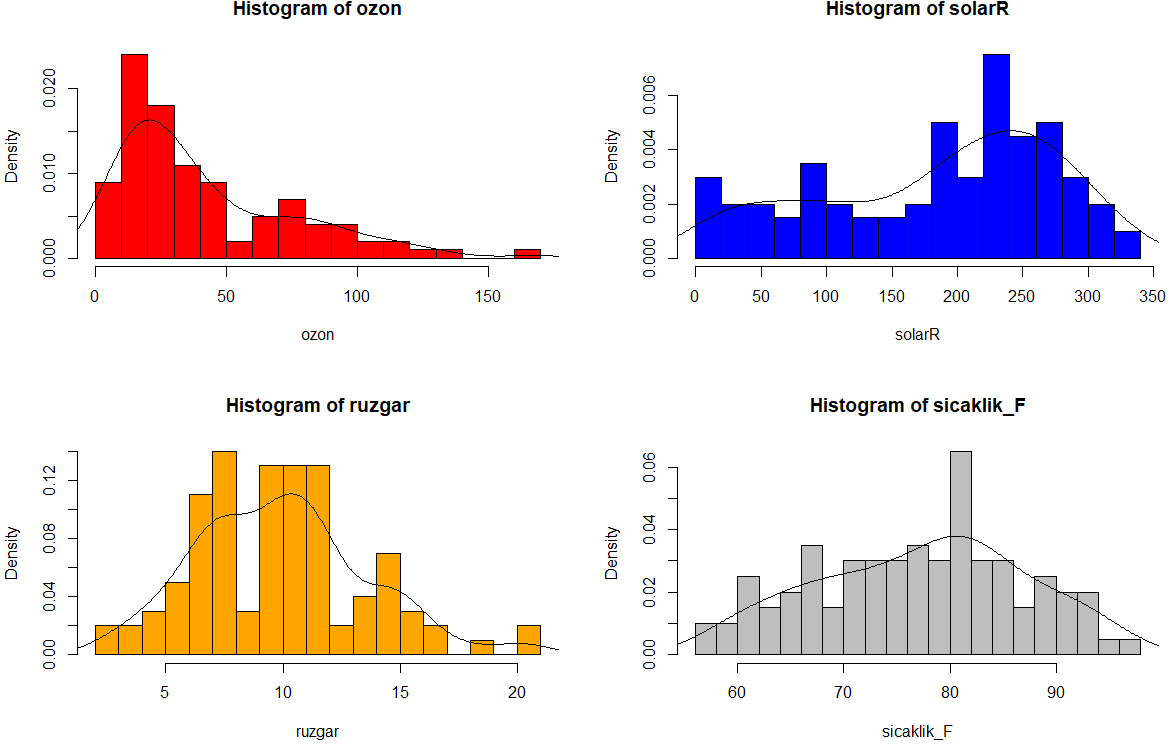


Burada ki en belirgin fark kullandığımız fonksiyonun ‘boxplot’ değil ‘Boxplot’ olması. İkisi arasında ki fark Boxplot grafiğinin x ve y koordinatlarına gerekli açıklamaları kendi ekliyor olması ve aykırı değerlerin ne olduğunu sayısal olarak yazıyor olması. Boxplot grafiği ‘Rcmdr’ (R Commander) paketinin içinde bulunan bir fonksiyondur ve bu fonksiyon hakkında ileride ayrıntılı bilgi verilecek.

Boxplot grafiği bize 1. Çeyreklik, medyan, 3.çeyreklik, en büyük ve en küçük değeri grafik üzerinde göstermektedir. Örnek vermek gerekirse Ozon miktarının aylara göre grafiğinde ağustos ayını ele alırsak medyan değerinin hemen hemen 50’ye eşit olduğunu ve tüm aylar içerisinde toplam 4 tane aykırı değer olduğunu söyleyebiliriz. Grafiklerin üst uç ve alt uç değerleri aykırı değerler yok kabul edilerek en büyük ve en küçük değerleri göstermektedir.

Şimdi 4 değişken için histogram grafikleri üzerinden yoğunluk tahminleri oluşturup bunların çizgilerini grafik üzerinde görelim.

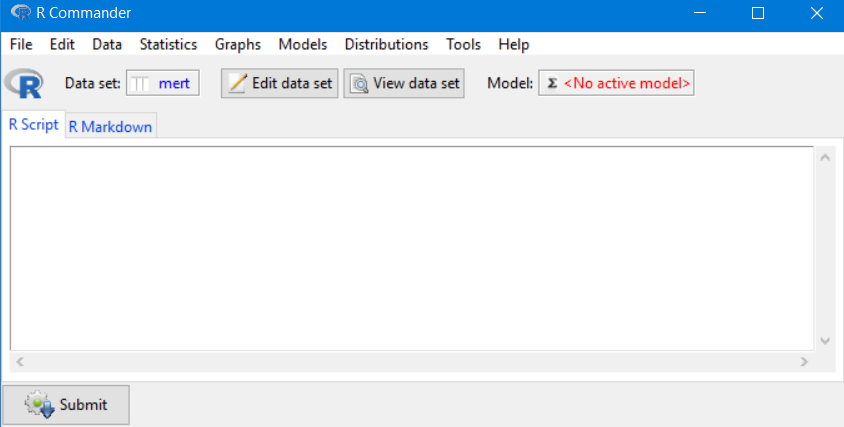




Görüldüğü üzere normal dağılım grafiğine yani çan eğrisi grafiğine yakın grafiklerin rüzgar ve sıcaklık değişkenleri olduğunu söyleyebiliriz.

**R Commander**

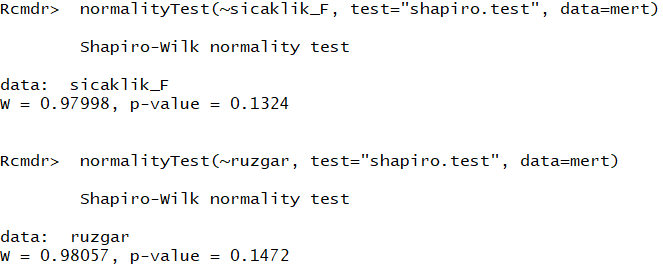
R Commander, basitçe anlatmak gerekirse istatistiksel her türlü analizi tek tuşla yapabildiğimiz ve aynı zamanda yazılan kodları bize vererek kendimizi geliştirmemize olanak sağlayan bir pakettir. R Commander’ın kurulum kısmını boxplot grafiklerini çizerken zaten yapmıştık. Güçlü bir paket olduğunu söylemekte fayda var yani yükleme kısmı biraz uzun sürebilir. library(Rcmdr) kodunu çalıştırdığımız anda program karşımıza gelecektir.



Programın görünüşü bu şekildedir. Kendi veri setimizi programa tanıttıktan sonra yukarıda statistics ve graphs kısımları içerisinde doğrudan istatistiksel analizler yapılabilmektedir. Şimdi bu paketi kullanarak analiz yapmaya devam edelim.

Yukarıda oluşturduğumuz histogram grafiklerinden rüzgar ve sıcaklık değişkenlerinin normal dağıldığı yorumunu yapmıştık şimdi bunu R Commander içerisinde ki normallik testini kullanarak doğrulayalım.

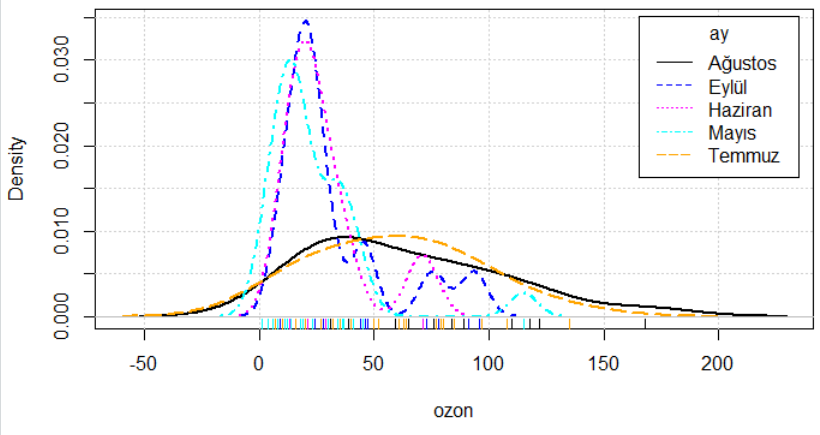




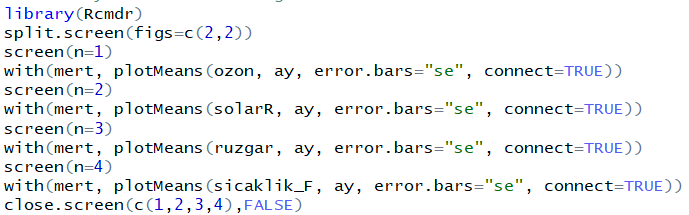
Görüldüğü üzere p-value değerleri Ho hipotezini kabul etmektedir. Yani iki değişkeninde normal dağılım gösterdiği söylenebilir.

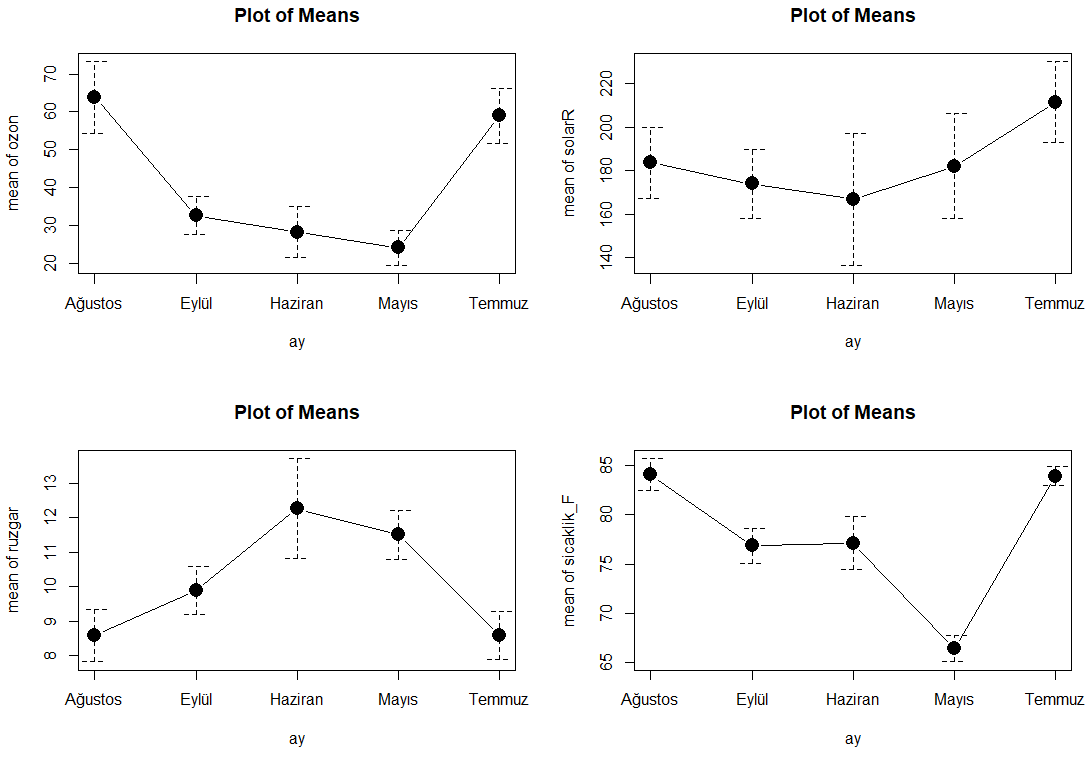
Şimdi, daha önceden çizdiğimiz yoğunluk tahmini grafiğinin çok daha güzel bir halini R Commander üzerinden çizdirelim.





Bu yol sayesinde her bir ay üzerinden her bir değişkenin yoğunluk tahminlerini görebiliriz. Diğer bir grafik kategorik değişkenin her bir elemanına göre ortalama üzerinden çizgi çeken fonksiyon.

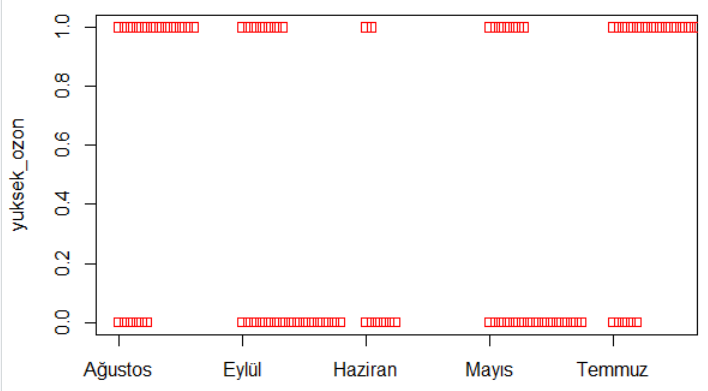




Görüldüğü üzere her bir değişkenin hangi aylarda artış gösterdiğini grafik üzerinden net bir şekilde görebiliyoruz.

Şimdi bize verilen veri setinde ki 4 ana değişkenin ortalamasından büyük ve küçük olarak oluşturulmuş ‘Yüksek’ değişkenlerini yorumlayalım. Bunu yaparken R Commander’ın içinde barındırdığı Strip Chart(Şerit Grafiği)’ı kullanalım.





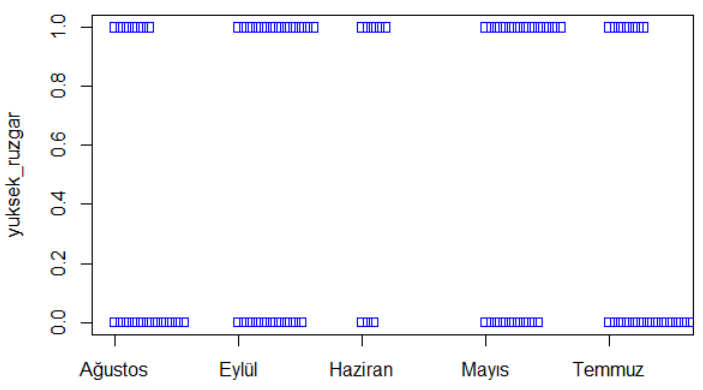
Yüksek ozon grafiği aylara göre yorumlandığı zaman ağustos ve temmuz aylarında ozon miktarının ortancanın üstünde seyrettiği yorumunu yapabiliriz. Değişkeni kategorik verilere göre ayırmadan genel olarak ortancadan düşük mü yoksa yüksek mi olduğunu araştıralım.



Bu 5 ayın genelinde ozon miktarının çoğunluğunun ortancadan düşük olduğunu söyleyebiliriz.

Diğer 3 veri içinde aynı şeyleri yapalım.

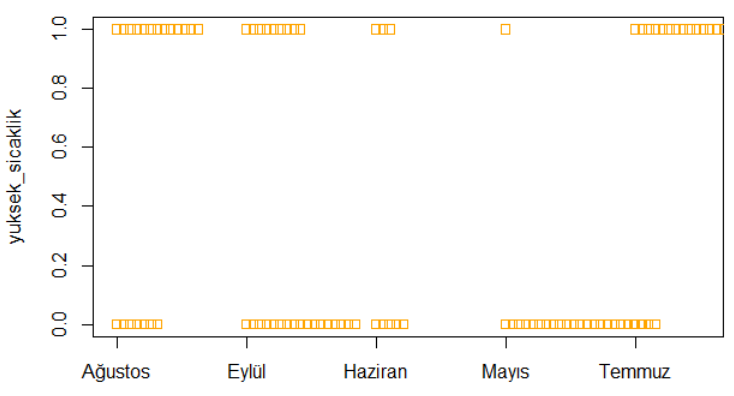






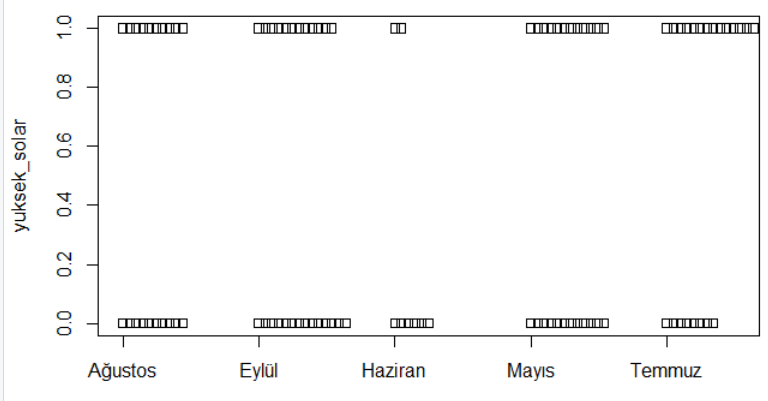
Yüksek Rüzgar verimizin de büyük çoğunluğu ortancadan düşük çıktı. Ortancadan yüksek olan ayların grafiğe bakarak eylül ve mayıs ayları olduğunu söyleyebiliriz.







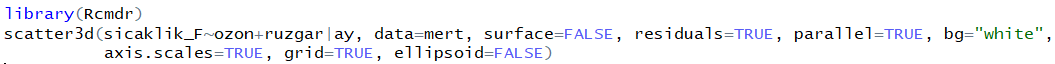
Yüksek sıcaklık verimizin de aynı şekilde büyük çoğunluğu ortancadan düşük çıktı. Her ne kadar mayıs ve temmuz biraz iç içe geçmiş olsa da ortancadan yüksek çıkan aylar ağustos ve temmuz ayları olarak yorumlanabilir. Kalan son değişkeni çizdirelim.

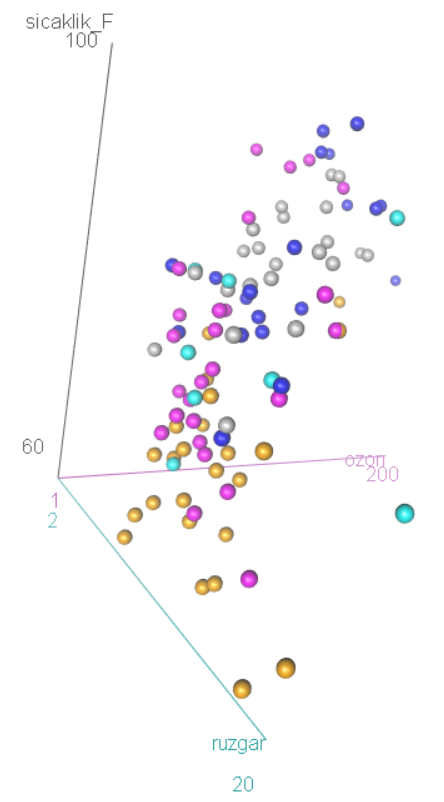




Burada verilerimizin büyük çoğunluğu diğer 3 değişkende olduğu gibi ortancadan düşük çıktı. Ortancadan yüksek çıkan tek ay ise temmuz olarak grafikten yorumlanabilir. Genel sonuç bütün değişkenlerin ortancadan düşük olduğudur dolayısıyla 5 ay içerisinde ortancadan yüksek olan ayların sayısının en fazla 2 olması beklenen bir sonuçtur.

Şimdi R Commander kullanarak elimizde ki rasgele 3 değişkenin 3 Boyutlu grafiğini çizdirelim.





Bu 3 boyutlu grafik üzerinde istediğimiz gibi yön çevirme işlemi yaparak analizi kolaylaştırabiliriyoruz. Grafik; rüzgar, ozon ve sıcaklık için çizdirilmiştir ve ay kategorik değişkenine göre renklendirilmiştir.

Şimdi özet kısmına gelelim.

**Özet:**

Genel olarak özetleyecek olursak elimizde ki 4 ana veri içerisinde ozon, solar radrasyon ve sıcaklık arasında pozitif bir ilişki olduğunu rüzgarın ise bunlarla negatif bir ilişki içinde olduğunu söyleyebiliriz. Buradan çıkan yorum rüzgarın bizim dostumuz olduğudur.

Saçılım grafiklerinden yaz aylarında sıcaklığın artmasıyla beraber ozon oranında artış ve aynı oranda rüzgar hızında azalış olmuştur.

Yoğunluk tahminlerinden rüzgar ve sıcaklığın normal dağılım, ozon ve solar radrasyonun normal dağılıma uymadığı görülmektedir.

Yüksek değişkenlerine bağlı analize baktığımız zaman 4 ana değişkenin hepsinin medyan yani ortanca değerinin altında sonuçlar gösterdiği yorumunu yapabiliriz. Bazı aylarda ortancadan yüksek çıkmıştır fakat genel çoğunluk alt değerlerdir.

Analiz de R Commander’ın kullanılmış olmasının çok yararlı olduğunu düşünüyorum. Bu paket program hem analizi daha güzel bir hale getirmiştir hem de yazdığı kod bloklarına bakarak neyin nasıl yapıldığını daha iyi anlamış ve öğrenmiş oluyoruz. Benim kendi yaptığım grafiklerin eksiklerini bu paket programın biraz da olsa kapattığını düşünüyorum.

**Mert Yanık**

**21522134**