**III. Aşama - Verinin Karakteristiklerinin Analizi:** İstatistik ve olasılık, sayısal analiz ve özellikle veri madenciliği dersi kapsamında elde ettiğiniz kazanımları burada değerlendirmenizi bekliyoruz. Veri seti içerisinde yer alan özelliklerin alt/üst aralıkları, kritik seviyelerin neler olabileceği, günün zaman dilimine göre, çalışma günlerine göre, tatil günlerine göre veya mevsimsel olarak nasıl bir değişime sahip olduğu ayrıca özelliklerin birbirleri arasındaki ilişkilerin olup olmadığının analizi beklenmektedir.

**Tanım:** Madenlerin hava kirliliği üzerinde etkisini daha iyi anlayabilmek adına çeşitli analizler yapılması gerekiyor. Bu analizleri yaparken de çeşitli bilim dallarından faydalanacağız. Bunlar istatistik bilimi, veri madenciliği gibi veri üzerinde analiz yapma imkânı sağlayan bilim dallarıdır.

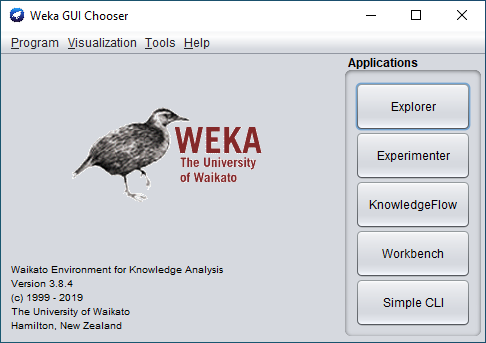
**Adım1:** Veri Madenciliğini Anlamak

**2.Veri Madenciliği Nedir?**

Dijital verinin toplanması ve saklanmasındaki gelişmeler, saklanan verilerin üstel bir şekilde büyümesine sebep olmuştur. Diğer yandan hayatın hızla elektronikleşiyor olması, internetin günlük yaşamın ayrılmaz bir parçası haline gelmesi ve teknolojinin alınabilir bir meta olarak toplum nezdinde yaygınlaşması toplanan veri artışını hızlandırmıştır.

Toplanan verilerin çeşidi her gün artmaktadır. Veri madenciliğinin büyük miktarda verinin anlamlı örüntüler bulmak amacıyla otomatik veya yarı otomatik yöntemler ile işlenmesi olduğunu söyleyebiliriz. Veri madenciliğinin en çok beslendiği bilimsel disiplinlerin istatistik ve makina öğrenmesi olduğunu söyleyebiliriz. Veri madenciliği disiplinler arası bir çalışmadır. İstatistik, veri tabanı teknolojileri, makina öğrenmesi, yapay zekâ ve görselleştirme gibi birçok farklı disiplin bünyesinde gelişen yöntemleri kullanır. Bahsi geçen disiplinler arasında sınırlar çizmek zor olduğu gibi, veri madenciliği ile bu disiplinler arasında da sınır çizmek zordur.

**Adım2:** Bizde bu projede çeşitli sonuçlar ve analizler elde etmek için veri madenciliği algoritmalarından birisi olan apriori ve C 4.5 algoritmalarını projemize uyarladık.



Bu algoritmaları hayata geçirebilmek ve sonuçlar üretip analizler yapabilmek için weka adlı program kullanılmıştır. Bu programdan elde edilen veriler bir sonraki hafta raporunda sonuç olarak sunulacaktır.

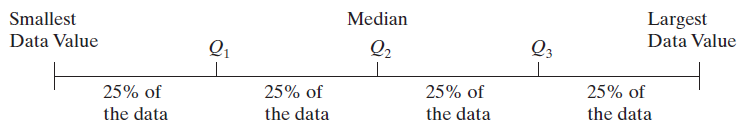
**Adım3:** Uç değer Hesabı (İstatistik Biliminden Yararlanıldı)

Uç değer hesabı verilerin dağılımlarını anlayabilmek ve yorumlamalar yapabilmek adına çok önemlidir. Uç değerler genel anlamda, veri setinde bulunan diğer değerler tarafından ifade edilen eğilimden büyük ölçüde farklı olan veri noktalarıdır. Başka bir deyişle, uç değerler diğer değerlerin dışında kalan değerlerdir. Bunları tespit etmek için çeşitli yazılımlar ve yöntemler olduğu gibi, veri tablolarında ve grafiklere bakarak ta anlaşılabilirler.

**Uç Değer Hesabı**

[Uç Değerler](http://mustafaakca.com/tag/uc-degerler/) Nasıl Belirlenir? Ekstrem bir değeri belirlemenin en kolay yolu şudur.

Ortalamalar veriyi ikiye bölen değerdir, Kartiller ise 4 bölen değerlerdir.

[](http://mustafaakca.com/wp-content/uploads/2014/08/kartil.png)

2 numaralı [kartil](http://mustafaakca.com/tag/kartil/) aynı zamanda ortalamadır. Çeyrekler arası mesafeyi kontrol etmek için şu formül kullanılır.

IQR=Q3-Q1

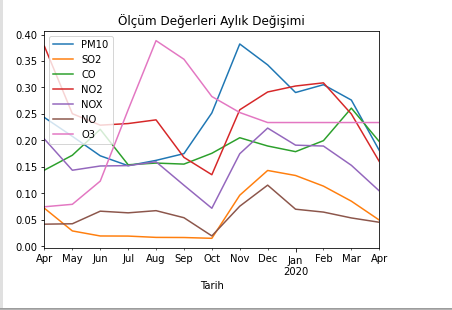
ve bir değerin [outlier](http://mustafaakca.com/tag/outlier/) ([uç değer](http://mustafaakca.com/tag/uc-deger/)) olup olmadığına ise şu şekilde bakılır.

Alt [UçDeğer](http://mustafaakca.com/tag/uc-deger/)  = Q1 – 1.5(IQR)  
Üst [Uç Değer](http://mustafaakca.com/tag/uc-deger/) = Q3 + 1.5(IQR)

Eğer bu değerleri aşan bir değer varsa, [outlier](http://mustafaakca.com/tag/outlier/)([uç değer](http://mustafaakca.com/tag/uc-deger/))’dir diyebiliriz.

Bu değerler Boxplot diyagramlarını hazırlamakta bize gerekli olacak değerlerdir. Q1, Q2([Ortalama](http://mustafaakca.com/tag/ortalama/)) ve Q3’ü bilinen bir dağılımın boxplot diyagramı alttaki gibi çizilebilir.

**NOT:**



3.Hafta raporunda kullanılan bu şekle bakılırsa veri setinde bulunan her bir verinin (ölçüm değerlerinin) uç değerleri net bir şekilde görünüyor.

Lakin bu uç değerler yukarıda verilen yöntemle ayrıyeten hesaplanarak projeye dâhil edilmiştir.