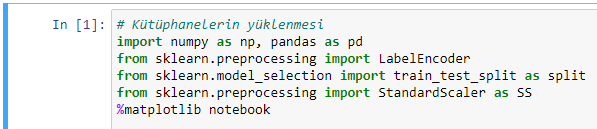
**2.Aşama – Verinin Görselleştirilmesi ve Sunumu**

Konya ilinin, hava kalitesini ölçmek amaçlı oluşturulan veri setini görselleştirmek için **“Anaconda”** programı kuruldu ve **“Notebook”** editörü kullanıldı.

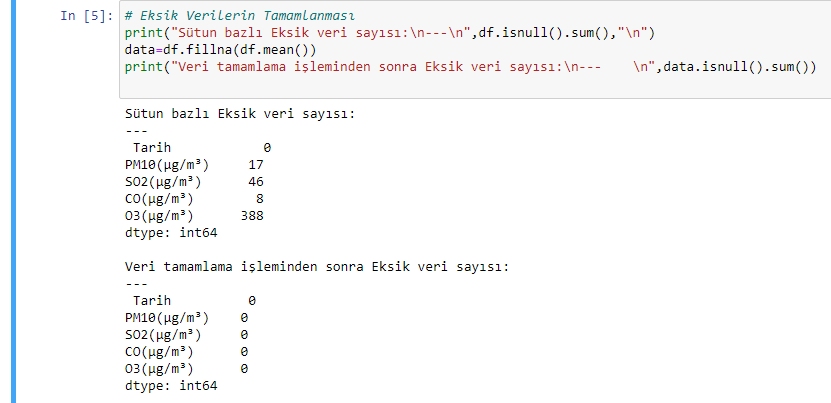
****Projede kullanılacak kütüphaneler, import edildi.

* Pandas: Veri tabanından veri okunabilir ve yazılabilir.
* Numpy: Diziler üzerinde işlem yapılabilir.
* LabelEncoder: Etiket oluşturur. Sınıflandırma yapmak için kullanılır.

**C:\Users\16541\Desktop\FİNAL_ymgk\Screen\3.Veri setinin yuklenmesi.PNG** Verisetinin yüklenmesi sağlandı. Veri setindeki sütun isimleri belirtildi. Ve ilk iki sütun, sütun isimleri olduğundan devredışı bırakmak için kodlar yazıldı.

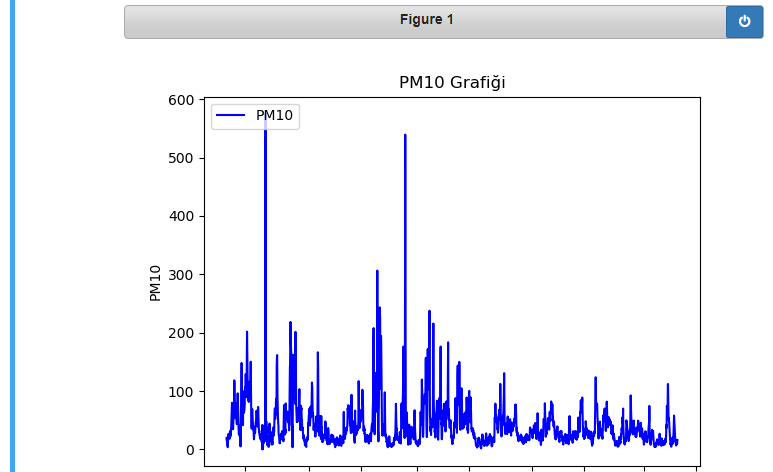
Veri bilimi projeleri, verinin keşfedilmesi ve temizlenmesi ile başlar ve bu işlemler projelerin en çok zaman alan kısımlarıdır. Dolayısıyla verinin keşfi ve temizlenmesi sırasında işleri kolaylaştıracak bir takım kütüphanelere ihtiyaç duyulur. Numpy verilerle çalışmayı oldukça kolaylaştırır. Numpy’ ın eksik kaldığı kısımlarda ise Pandas kullanılır. Ancak Pandas Numpy’ ın bir alternatifi olarak değil, uzantısı olarak düşünülebilir. Pandas, Numpy’ın sütun adları ve homojen olmayan verilerle çalışamama gibi eksik kaldığı kısımlara ve daha fazlasına çözümler üretir.

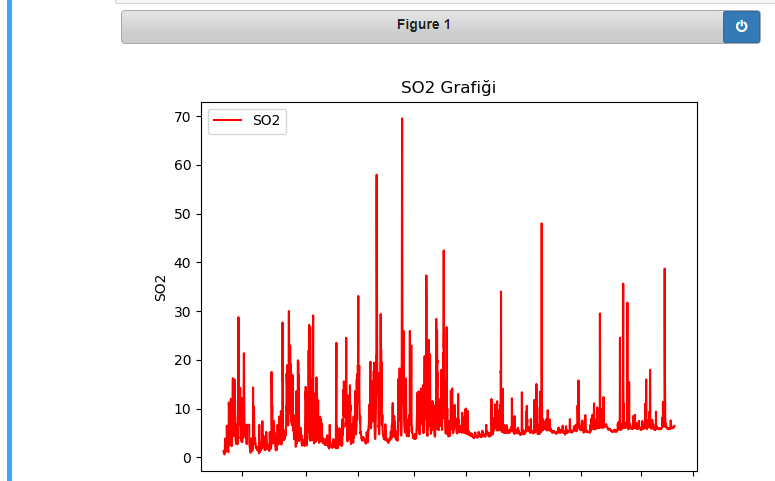
Veri setinde eksik değerler bulunmaktadır. Eksik verilerin tamamlanması için değişik yöntemler ortaya atılmıştır.  Genelde verinin hatalı okunması, veri kaynağında yaşanan bozulma,  bazı verilere erişim zorluğu gibi sorunlar eksik verilere sebep olabilmektedir. Birçok yöntem mevcut olmak beraber projede **Mean Substitution (yerine ortalamayı koyma) yöntemi kullanıldı.** Bu yöntem, veri setinde kayıp verinin olduğu alandaki diğer verilerin ortalamasını alarak kayıp olan verilerin yerine bu değeri yazarak  doldurmaya yarayan yöntemdir. Veri aralığı düşük olan verilerde kullanıldığında yararlı olabilir. Aksi halde çok değişkenli veri setinde  hata oranını arttırır. Bu proje de sütunda bulunan maksimum ve minimum değerlerine bakıldı ve değerler birbirine yakın olduğu tespit edildi. Çok büyük problem yaratılacağı düşünülmediği için bu yöntem kullanılabilir olduğuna karar verildi.

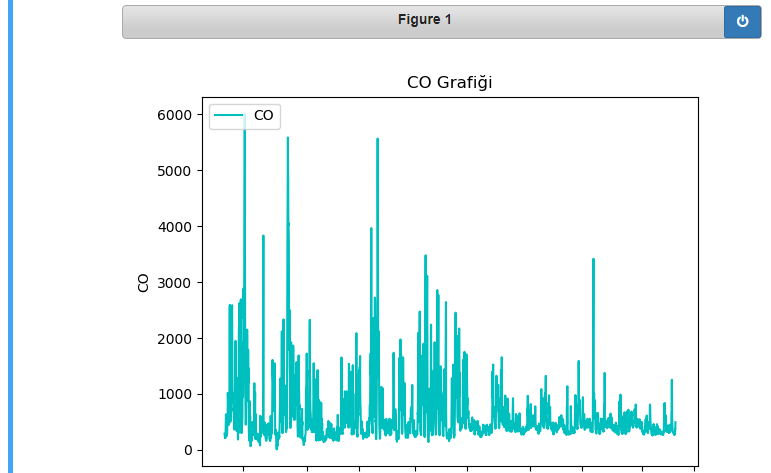
****

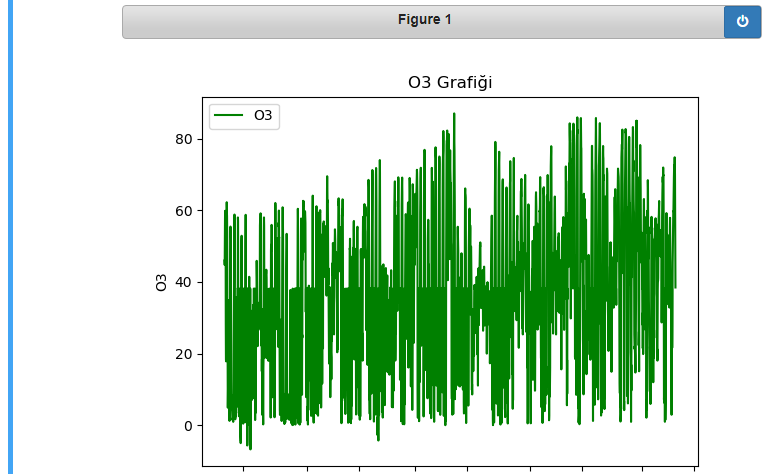
Veri setindeki PM10 sütununda 17 eksik veri olduğu tespit edildi. Uygulanan yöntem ile PM10 sütununda bulunan bütün değerlerin ortalaması alındı ve bu değer, boş olan 17 hücreye atandı. Aynı işlemler diğer sütunlara da uygulandı. Artık veri setinde eksik veri bulunmamaktadır.

Görselleştirme işlemi yapıldı.











Görselleştirme matplotlib ile yapıldı. Her bir özelliğin tarihe göre değişimi gösterilmiştir.