# GÜNCEL VERİ TEMİZLEME ARAÇLARI

YZ5521, VERİ ANALİTİĞİ

Yusuf CAN
yusufcan5@mu.edu.tr

11 Kasım 2021

**Özet**

Günümüzde şirketler ve kuruluşlar olağanüstü miktarda veri üzerinde işlemler yapmaktadır. Yalnızca on yıl öncesine kadar bir gigabayt veri büyük miktarda bir veri iken şimdilerde kimi kuruluşlar zetabaytlar seviyesinde büyük hacimli verileri yönetmektedir.

Veri bilimi, kısaca verilerden anlamlı sonuçlar çıkarma süreçleridir. Ancak verilerden anlamlı sonuçlar çıkarabilmek için temiz veriye gereksinim duyulmaktadır. Bu noktada veri bilimcileri veya veri analizcileri gibi veriyle uğraşanların veriyi incelemeye hazır hale getirebilmeleri için veri temizleme araçlarına ihtiyaçları doğmaktadır. Veri temizlemek ve analiz etmek için veriyle uğraşanların kullanımına sunulmuş manuel ve otomatik kullanım imkânı olan birçok veri temizleme aracı bulunmaktadır.

## 1. Giriş

Veri bilimi, birçok farklı alandan elde edilen verilerden yola çıkarak, bilgisayarların hesaplama gücü ve makine öğrenmesi gibi algoritmalar yardımı ile verilerden anlamlı sonuçlar çıkarıp bu verileri faydalı çıktılara dönüştürme süreçleridir. Bu süreçlerde verileri, temizleme, azaltma, dönüştürme, görselleştirme gibi işlemler yapılarak veri işlenilecek hale getirilerek tahmin edilen sonuçların doğruluğunun yüksek olması amaçlanır. Bu süreçteki işlemlerden birisi belki de en önemlisi veri temizleme işlemidir.

## 2. Veri Temizleme

Veri temizleme, veri tabanından eksik değerlerin kaldırılması, tutarsız verilerin düzeltilmesi, veri toplama ve gürültülü verilerle mücadele vb. gibi birkaç faktörü ekleyerek veri kalitesini artırma işlemidir. Veri temizleme, bir veri bilimci veya veri analizcisinin öğrenmesi ve bilmesi gereken öncelikli konulardan biridir. Günümüzde müşteri verileri veya sosyal medya verileri gibi alanlardan elde edilen olağanüstü miktarda veri bulunmaktadır. Yüksek miktardaki bu veri doğal olarak düzgün bir şekilde depolanamayabiliyor veya veriden çıkarımlar yapılmak istenildiğinde işlenilecek durumda veya hazır olmayabiliyor. Bu gibi durumlarda veri temizliği ön plana çıkmaktadır. Çünkü veri kaliteli değilse kalitesiz veri ile iyi sonuçlar elde etmek çok mümkün değildir.. Ancak sürekli artan yüksek miktardaki bu verileri incelemek, temizlemek, analiz etmek, görselleştirmek için birçok araç da artan veri ile doğru orantılı olarak artmaktadır. Veri temizlerken yapılması gereken işlemler şu şekilde sıralanabilir.

1. Çözmeye çalışılan problemle ilgili olmayan ve istenmeyen gözlemler veri kümesinden çıkarılmalıdır.

2. Veriler standartlaştırmalıdır. Yani veri kümesindeki verilerin aynı ölçüde olması sağlanmalıdır.

3. Farklı kaynaklardan gelen veriler bir ana yapıda birbirleriyle eşlenerek tutarlı olmaları sağlanmalıdır.

4. İstenmeyen aykırı değerler veri kümesinden çıkarılmalıdır. Aykırı değerler eğer hatalı değilse faydalı olabilirler, ancak hatalı olmaları durumunda sonuçların hatalı olmasına neden olurlar. Ayrıca hangi aykırı değerlerin kullanılacağı, hangilerinin kaldırılacağı konusunda da bir karar verilmesi gerekmektedir.

5. Farklı veri kaynaklarından gelen verilerin birbirleriyle çelişmemesi sağlanmalıdır.

6. Veri tiplerini dönüştürmek ve sözdizimi (syntax) hatalarını çözmek gerekmektedir. Bunlar, boşlukları kaldırma, yazım hatalarını kontrol etme ve düzeltme, verilerin doğru şekilde sınıflandırılmasını sağlama gibi işlemler olabilir. Örneğin sayısal veriler üzerinde çalışılıyorsa veri tipi tam sayı, ondalıklı sayı vb. olmalı ve ona göre etiketlenmelidir.

7. Eksik veriler var ise bunlar kaldırılabilir, diğer veriler ile olan ilişkilerinden istatiksel çıkarımlarda bulunarak doldurabilir ve etiketlenebilir.

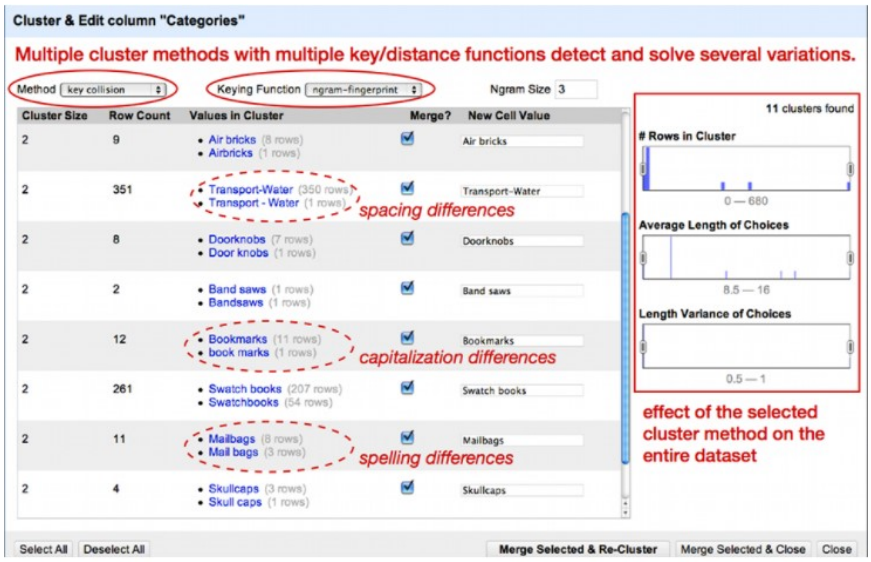
8. Verileri doğrulama bu sürecin son adımıdır. Veri temizleme sürecinde buraya kadar gerçekleştirilen adımların doğru şekilde yapılıp yapılmadığı kontrol edilir. Bu süreçte sık sık geriye dönüp daha önce gerçekleştirilen adımların tekrarlanması gerekebilir. [1]

Python, R dilleri ve MS Excel gibi araçların hepsi veri temizliği için çok önemli araçlar olarak ön plana çıksa da, giderek artan sayıda veri temizleme aracı mevcuttur. Verilerle çalışanlar için doğru veri temizleme aracının seçilmesi bu işin önemli parçasıdır.

## 2.1. Veri Temizleme Araçları

### 2.1.1. Open Refine

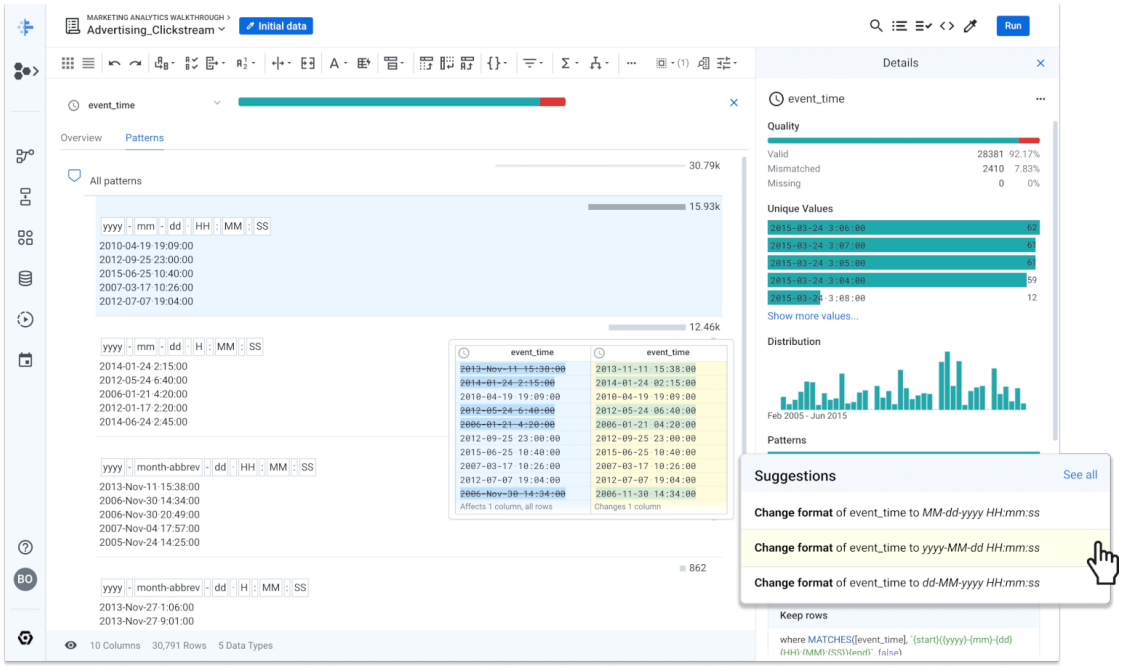
Daha önceden Google Refine olarak bilinen Open Refine, en çok bilinen açık kaynaklı veri temizleme aracıdır. En önemli avantajlarından birisi açık kaynak olduğu için ücretsiz olması ve özelleştirilebilmesidir. OpenRefine verileri farklı formatlar arasında dönüştürmeye ve verilerin temiz bir şekilde yapılandırıldığından emin olunmasına imkân sağlamaktadır. Excel gibi tablo yazılımlarına benzese de ilişkisel veri tabanı gibi davranır. Bu özelliği, basit Excel operasyonlarında daha ziyade biraz daha karmaşık işlemler yapan veri analizcileri için kullanışlı bir araç olmasını sağlamaktadır. Diğer bir önemli avantajı, makinenizdeki verilerle çalışabilme imkânı sunarak güvenlidir. Ayrıca OpenRefine'i harici web hizmetlerine ve buluttaki diğer kaynaklara bağlayarak veri kümeleri üzerinde işlemler yapılabilir. Gerektiğinde veriler merkezi bir veri tabanına da yüklenebilir. 15'ten fazla dil desteği bulunmaktadır. [2]




Şekil 1: Open Refine Veri Temizleme Aracı [3]

### 2.1.2. Trifacta Wrangler

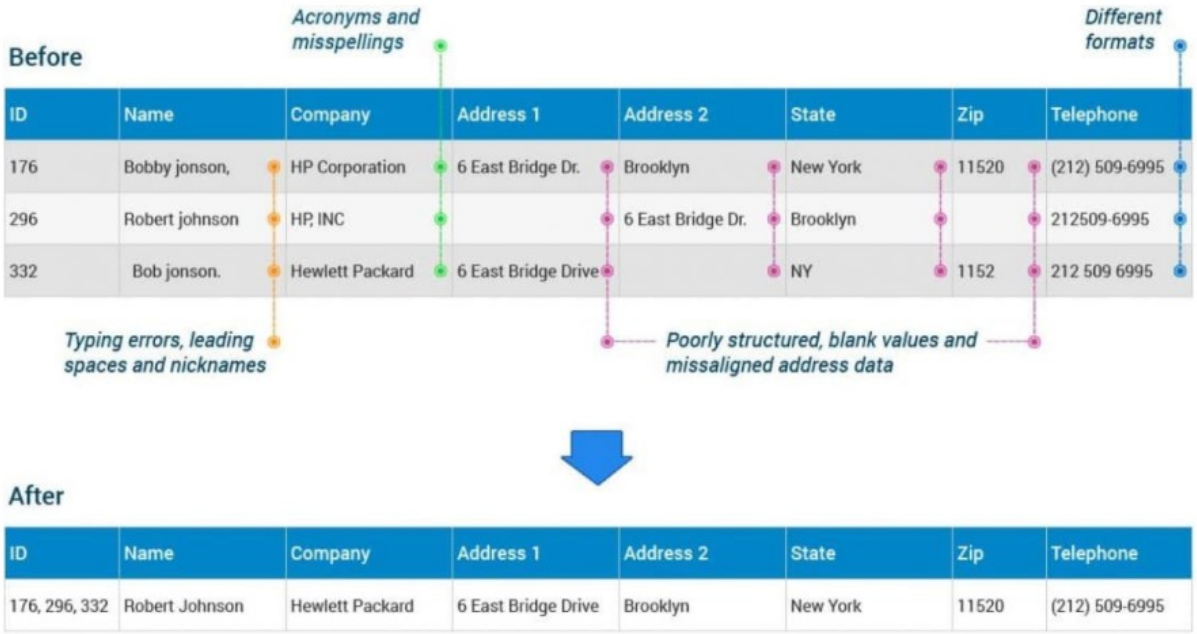
Trifacta Wrangler, verileri dönüştürmeye, analizler yapmaya ve görselleştirmeler oluşturmaya olanak tanır. Tutarsızlıkları tespit etmek ve önerilerde bulunmak için makine öğrenimi algoritmalarını kullanan bu araç, veri temizleme sürecini büyük ölçüde hızlandırmaktadır. Örneğin, yapay zekâ algoritmaları kullanılarak, aykırı değerler kolayca tanımlanabilir veya kaldırılabilir, genel veri kalitesini izleme otomatik hale getirilebilir. Ayrıca, sıfırdan veri pipeline'ı oluşturmak yerine (ciddi zaman alıcı bir iş yükü), aracın kullanıcı ara yüzü bunun çok daha görsel ve sezgisel bir şekilde yapılmasına imkân sağlamaktadır. Bu aracın sürümlerine göre kullanım imkânları değişmektedir. Örneğin, Wrangler Pro daha büyük veri kümelerini ve bulut depolamayı desteklerken giriş (enterprice) sürüm, ekipler halinde çalışmak için gerekli araçları sağlar. Ayrıca her verinin hassas veri olduğundan yola çıkarsak bu araç merkezi güvenlik yönetimine sahiptir.




Şekil 2: Trifacta Veri Temizleme Aracı [4]

### 2.1.3. Winpure Clean & Match

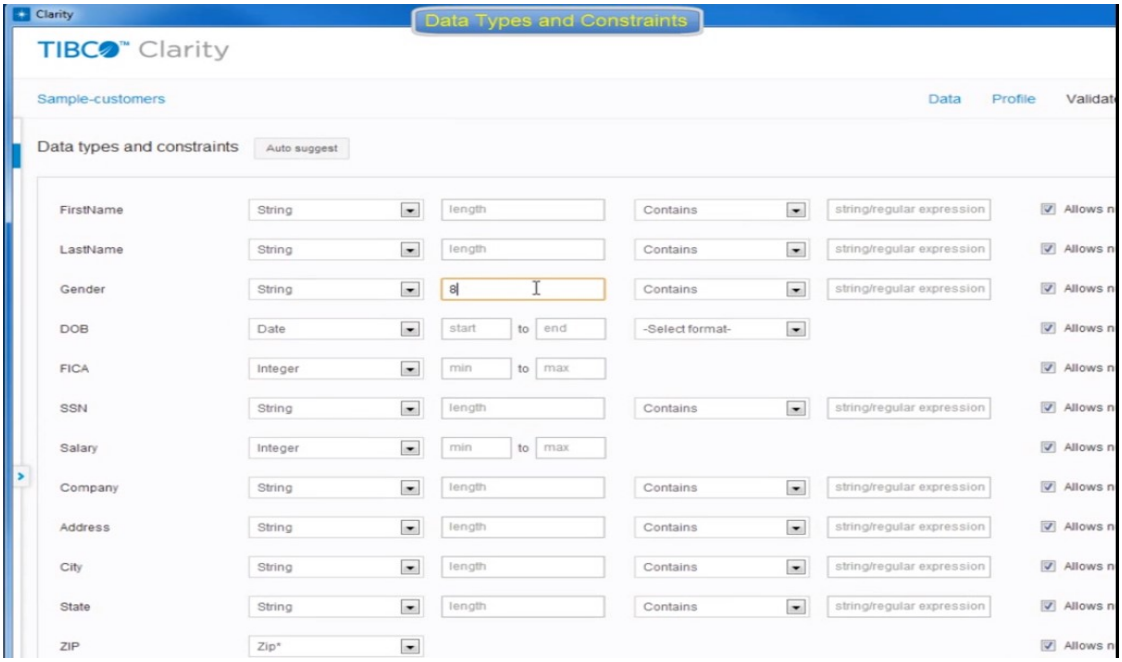
Trifacta Wrangler'a biraz benzeyen ödüllü Winpure Clean & Match, sezgisel kullanıcı ara yüzü sayesinde verilerin temizlenmesine, tekilleştirilmesine (drop dublicates) ve çapraz eşleştirmeye (cross-match) imkân sağlayan bir araçtır. Yerel (local) olarak kullanıma uygun olduğundan veri kümesi buluta yüklenmediği sürece veri güvenliği konusunda endişelenilmesine gerek yoktur. Bu, özellikle müşteri verileri (CRM verileri ve email adresleri) gibi hassasiyet gerektiren veriler üzerinde çalışanlar için oldukça önemli bir özelliktir. Ayrıca Winpure Clean & Match, birçok veri tabanı ve elektronik tablolarla birlikte çalışır. Diğer kullanışlı özellikleri arasında bulanık eşleştirmeye (eşleşmelerin rastgele kısaltmalara veya yazım hatalarına göre nerede farklılık gösterdiğini belirlemeyi içerir) ve kullanıcının programlayabileceği kural tabanlı temizleme prosedürlerinin oluşturulmasına olanak sağlar. Almanca, İngilizce, Portekizce ve İspanyolca olmak üzere 4 ayrı dilde kullanım imkânı vardır. Ücretsiz sürümünün çok sayıda özellik sunmasından dolayı küçük çaplı işletmeler için ideal bir seçenek haline gelmektedir.




Şekil 3: Winpure Clean & Match Veri Temizleme Aracı [5]

### 2.1.4. TIBCO Clarity

Bulut tabanlı bir yazılım hizmeti olan (SaaS) TIBCO Clarity, ham verileri temizlemek ve tümünü tek bir yerde analiz etmek için idealdir. XLS ve JSON dosyalarından sıkıştırılmış dosya formatlarına, çevrimiçi havuzlardan (repositories), veri ambarlarına kadar birçok farklı kaynaktan veri alan, zengin özelliklere sahip bir veri temizleme aracıdır. Bununla birlikte, TIBCO, veri eşleme, ayıklama, dönüştürme, yükleme (Extract Transform Load (ETL)), veri profili oluşturma, örnekleme ve toplu işlevsellik, tekilleştirme gibi birçok özellik sunmaktadır. Ayrıca, yapılan bir değişiklik geri alınmak istenildiğinde dönüşüm geri alma (transformation undo) gibi emsalleri arasında fark oluşturan ve sahip olunması gereken bazı yararlı özelliklere sahiptir. Bahsedilen onca özellik ve fonksiyonlarının yanında tek dezavantajı, ücretsiz bir sürümünün olmamasıdır, ancak TIBCO Clarity hala önemli ve sağlam bir yazılım olarak sektördeki yerini korumaktadır.




Şekil 4: TIBCO Clarity Veri Temizleme Aracı [6]

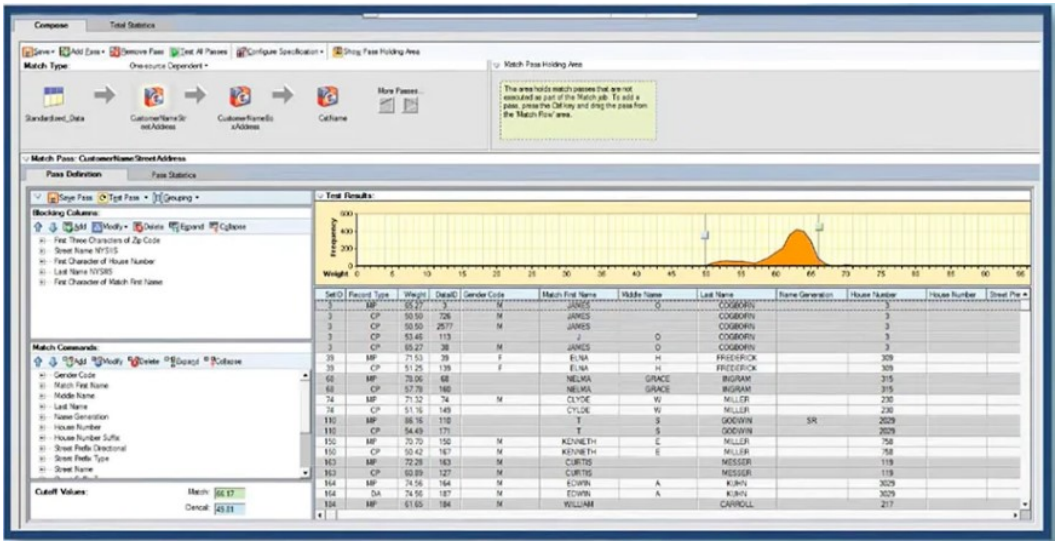
### 2.1.5. Melissa Clean Suite

Veri temizleme ve yönetim aracı olan Melissa Clean Suite, birçok işletmenin kullandığı Salesforce ve Microsoft Dynamics gibi müşteri ilişkileri yönetimi (CRM) sistemlerini desteklemek için özel olarak tasarlanmıştır. İlgi alanı bu iki sistem olduğu için tüm standart Salesforce nesnelerini destekler ve Dynamics'teki standart formlarla bütünleştirme imkânı vardır. Basit ve anlaşılır bir yapıya sahip olduğu için öğrenmesi kolaydır. Bunlar, demografik oluşturma, veri hedefleme ve segmentasyon içerir. Melissa Clean Suite'in ana avantajı, verilerin toplanırken temizlemesidir. Bu özellik, sonradan harcanacak çabayı minimuma indirir. Örneğin, kişileri sisteme girmeden önce otomatik olarak tamamlar, düzeltir ve doğrular. Veri sisteme girdikten sonra gerçek zamanlı temizleme ve toplu işleme ile proaktif olarak veri kalitesini korur. Pazarlamayla ilgili veri etkinliklerini hedef almasına rağmen, Melissa'nın genel veri yönetimi perspektifinden de zaman kazandıran net faydaları vardır.


Şekil 5: Melissa Clean Suite Veri Temizleme Aracı [7]

### 2.1.6. IBM Infosphere Quality Stage

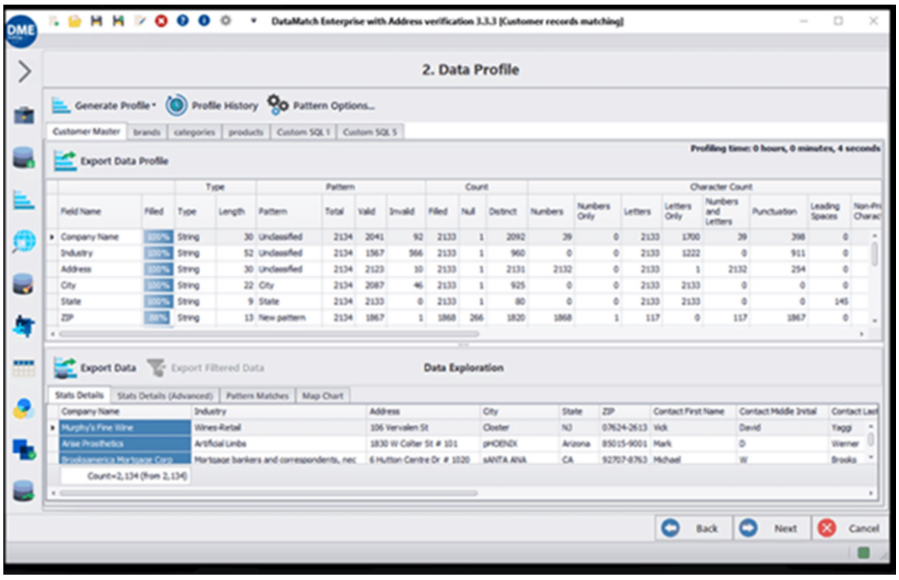
IBM Infoosphere Quality Stage, IBM'in sunduğu veri yönetimi araçlarından biridir. Adından da anlaşılacağı üzere veri kalitesi ve yönetimine odaklanır. Veri eşleştirme, tekilleştirme vb. ile ilgilenirken özellikle büyük verileri (big data) temizlemek için tasarlanmıştır. Bu amaçla, yaklaşık 200 yerleşik veri kalitesi kuralına sahiptir ve bu özellik sayesinde analizcilere ciddi manada zaman kazandırır. Bunların yanında temel özelliklerinin tümü, veri ambarı, veri yönetimi ve geçiş (migration) gibi yoğun emek gerektiren işlemleri destekler. Kurum içi veya bulutta kullanılabilen bu araç, aynı zamanda veri profili oluşturma imkânı sunmaktadır. Geniş bir veri tabanı görünümünden verilerin içeriğini, kalitesini ve yapısını keşfetmek veya sütunları ayrı ayrı analiz ederek detaylı incelemeler/ analizler yapmak için kullanılabilmektedir. Teknik bilgisi olmayanlar için en iyi araç olmasa da, veri kalitesi anlamında yüksek puanlara (score) sahiptir. Temizlik bakımından veri kalitesinin yüksek olması teknik yetenekten bağımsız olarak herhangi bir kullanıcının bir veri kümesinin bütünlüğü hakkında genel bir fikir edinmesini sağlar. Bu özellik, yönetici pozisyonundaki personel için çok önemli bir özelliktir.



Şekil 6: IBM Infoosphere Quality Stage Veri Temizleme Aracı [8]

### 2.1.7. Data Ladder (Datamatch Enterprise)

Datamatch Enterprise, Data Ladder tarafından görsel olarak yönlendirilen bir veri temizleme uygulamasıdır. Yukarıda bahsedilen diğer araçlar gibi müşteri verilerine odaklanır. Ancak diğerlerinden farklı olarak veri kalitesi bakımından düşük puanlara sahip verilerin kalitesini yükseltmek ve bu sorunları çözmek için tasarlanmıştır. Kullanımı basit olan bu araç, baştan sona veri işleme süreci boyunca kullanıcıyı desteklemek için gözden geçirme ara yüzü kullanır. Çok çeşitli içe ve dışa aktarma işlevlerini kullanarak, karmaşık dâhili iş prosedürleriyle uyumlu veri tabanı tablolarından Excel elektronik tablolarına kadar birçok işlemi yapabilme imkânı sunar. Ayrıca kullanıcıların büyük ve küçük veri kümelerinde veri tekilleştirmesine, ayıklamasına, standartlaştırmasına ve veri eşleştirmesine olanak tanır. Hedeflenen sonucun ne olduğuna bağlı olarak, doğruluk söz konusu olduğunda çeşitli güvenlik seviyelerine yanıt vermek için eşleşme tanımlarını manuel olarak yapılandırılabilir. İşlemler önceden zamanlanarak, veri temizleme görevleri önceden ayarlanabilir ve programlanabilir. Veri temizlemenin bir kerelik bir iş olmadığını bir süreç olduğunu dikkate alacak olursak bu özelliğiyle Datamatch Enterprise ön plana çıkmaktadır.




Şekil 7: Data Ladder (Datamatch Enterprise) Veri Temizleme Aracı [9]

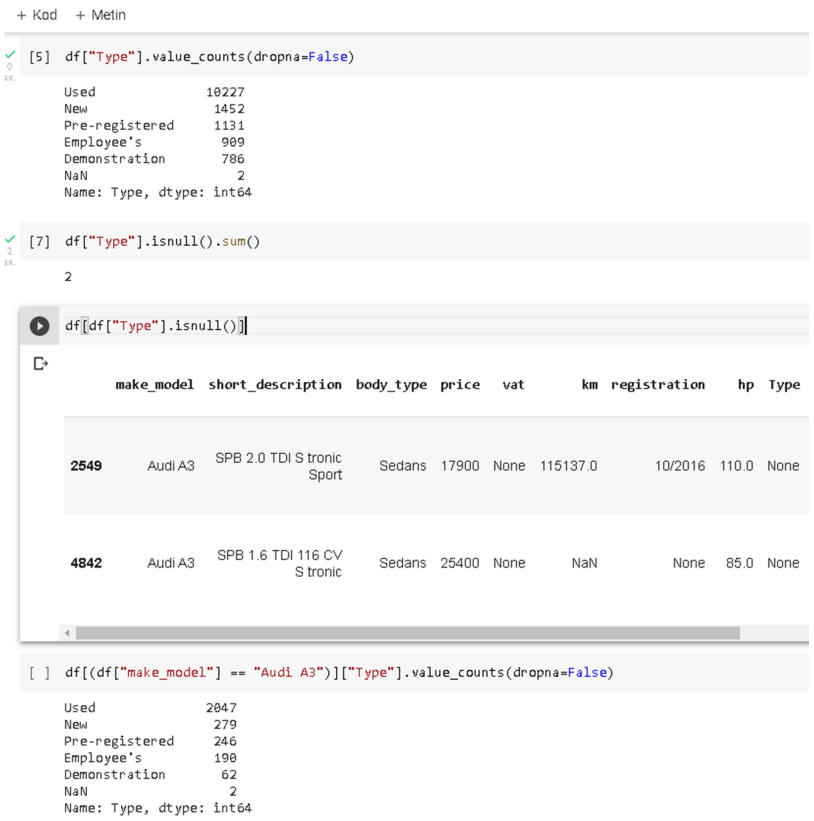
### 2.1.8. Python

Pandas, dataframe diye tabir ettiğimiz yapılar için oluşturulmuş, veri analizi ve veri ön işlemeyi kolaylaştıran açık kaynak kodlu bir kütüphanedir. [10] Pandas dağıtık işlemeye uygun olmadığı için işlenilecek verinin büyüklüğü makinenin kapasitesiyle sınırlıdır. Ancak dataframe'ler üzerinde işlemler yaparken hızlı ve etkili bir kullanım sağlar. Dosyalar arası geçişler çok kolaydır. Csv ve text dosyaları incelenebilir sonuçlar bu dosya tiplerine kolay bir şekilde kaydedilebilir. Kayıp/ eksik veriler (missing datas) üzerinde işlemlerin yapılabilmesini kolaylaştırır. Zaman serisi (time series) gibi özel veri kümelerinin analizinde oldukça yardımcı bir kütüphanedir. Ayrıca özel fonksiyonları sayesinde (reshape gibi) veri daha etkili bir şekilde kullanılabilir.

Numpy, Python'da bilimsel hesaplamalarda kullanılan temel pakettir. [11] Çok boyutlu diziler (array), çeşitli türetilmiş nesneler (maskelenmiş diziler ve matrisler gibi) ve birçok matematiksel, mantıksal, şekil manipülasyonu, sıralama, seçme gibi diziler üzerinde hızlı işlemler yapılmasını sağlayan Python kütüphanesidir. Numpy kullanılarak istatistik işlemler ve simülasyonlar da yapılabilir. Numpy dizileri çok sayıda veri üzerinde gelişmiş matematiksel işlemleri ve diğer işlemleri kolaylaştırır. Benzeri işlemleri, Numpy paketi kullanmadan yapmak daha zordur ve vakit alır.

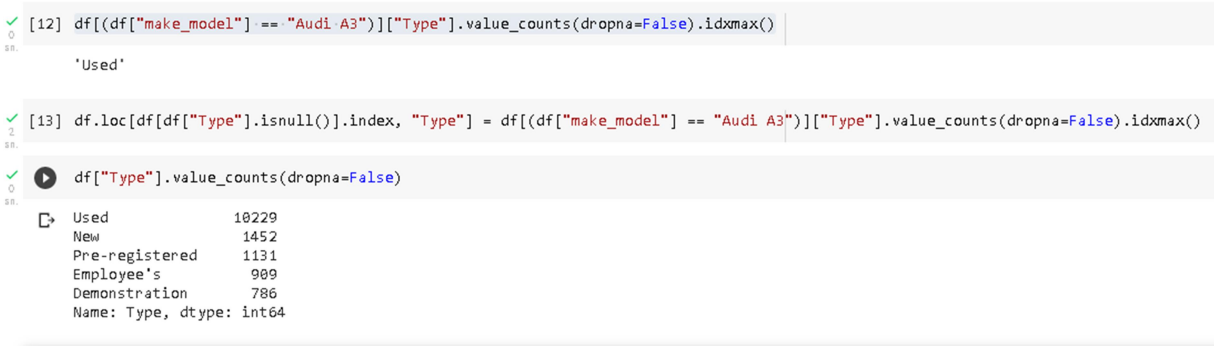
Veri temizleme işlemlerinde pandas ve numpy kütüphanelerinin kullanımı oldukça yaygındır. Bu kütüphaneleri kullanabilmek için Jupyter Notebook, Visual Studio, RStudio, PyCharm [12] gibi çeşitli tümleşik geliştirme ortamları (Integrated development environment IDE) mevcut olmakla beraber Google Colab gibi herhangi bir kurulum gerektirmeyen tamamen bulut üzerinde çalışan notebook ortamları da mevcuttur.

Örneğin Şekil 8'de gösterilen bir dataframe'de "Type" sütununda 2 eksik veri mevcuttur. Bu veriler, veriyle uğraşan kişinin sezgilerine ve tecrübesine bağlı olarak veri kümesinden çıkarılabilir veya genel istatiksel metotlarla doldurulabilir.




Şekil 8: Google Colab Ortamında Veri Analizi ve Temizleme

Şekil 9'da, "Type" sütunundaki eksik verilerin incelenmesi sonucu eksik olan "Type" verileri en çok tekrarlayan veri (Used) ile doldurulmuştur. Bu, eksik verileri doldurmada kullanılan yöntemlerden sadece birisidir.




Şekil 9: Google Colab Ortamında Veri Analizi ve Eksik Verilerin Doldurulması

## 3. Sonuç ve Öneriler

Farklı işlevlere ve kullanım alanlarına göre çok çeşitli veri temizleme aracı bulunmaktadır. Burada bahsedilen araçlar veri analizcileri ve bilimcilerinin günlük çalışmalarında sıklıkla kullandıkları veri temizleme araçlarındandır.

Yazılımlar veri temizleme işlemlerinde kullanıcılara geniş imkânlar sunsa da en iyi veri temizleme aracı veri ile uğraşanların kendi bilgi, tecrübe ve sezgileridir. Yazılımların çoğu süreci otomatikleştirse de daha karmaşık işlemlerin yüksek oranda manuel işlemler gerektirmektedir. Yazılımlar bu işlemleri yapmak için sadece birer araçtır. [13] Bu yüzden veri uğraşanların ve veri

bilimi alanında iş sahibi olmak isteyenlerin Excel, Python, R ve ilişkisel veri tabanları gibi araçları kullanarak manuel veri temizleme becerilerini sürekli geliştirmesi gerekmektedir.

## Kaynakça


1. https://careerfoundry.com/en/blog/data-analytics/best-data-cleaning-tools/

2. https://openrefine.org/

3. https://programminghistorian.org/en/lessons/cleaning-data-with-openrefine

4. https://www.trifacta.com/products/

5. https://winpure.com/products/clean-match/

6. https://clarity.cloud.tibco.com/landing/tutorial.html

7. https://appexchange.salesforce.com/appxListingDetail?listingId=a0N30000000pvskEAA

8. https://www.ibm.com/products/infosphere-qualitystage

9. https://dataladder.com/products/datamatch-enterprise/

10. https://pandas.pydata.org/

11. https://numpy.org/

12.https://businessoverbroadway.com/2020/07/14/most-popular-integrated-development-environments-ides-used-by-data-scientists/

13.https://careerfoundry.com/en/blog/data-analytics/data-wrangling/#what-tools-do-data-wranglers-use