

1- Checking Data Quality:

UNIQUE FONKSİYONU

*Veriler güvenilir ve değersiz arasında herhangi bir yere düşebilir. Çok fazla insan elde ettikleri verilerin güvenilir olduğunu varsayar ve dalış yapar. Analiz yapmaya ve ilham verici kararlar alırlar fakat sonucu yanıltıcı olabilir.

*Bu yüzden kendimize sormamız gereken ilk soru bu verilere güvenebilir miyiz? Neden? Neden bunu sormalıyız? Çünkü veri yönetimimizden ve veri analizimizden etkilenen insanlar vardır. Bu yanlış verilere maruz kalmış birçok insan olabilir.

*Verinin kalitesini kontrol etmeliyiz.

**Örn: Bir excel tablosunda sınavda geçme/kalma durumu varsa filtrelerden bunu kontrol etmeliyiz. Geçme(P) ve kalma(F) dışında herhangi bir durum varsa onu ortadan kaldırmamız gerekir. Böylece verimiz daha güvenilir hale gelir.

*Eğer bir sütunda çok fazla kısaltma varsa K.C., K C, KC, K.C gibi aslında bunların aynı kişiye veya nesneye ait olduğunu sadece kısaltmalarının yanlış kullanıldığını görebiliriz. -

*Excel'de herhangi bir boşluğa =unique(tüm sütunu seç) sonrasında enter yaparsak sütundaki kişileri nesneleri tek seferde görebiliriz ve buradan da hangi verinin çok kez yanlış yazıldığını görürüz.

*İsimleri tek tek bulup elde değiştirmek yerine şunları yapabiliriz:

-Bul ve .seç

-Değiştir

-Tümüne uygula

PIVOT TABLE

*Örneğin gelen verileri çeşitli şekillerde analiz edip her temsilcinin satış tutarlarını kontrol etmek istiyoruz. Ürün hatlarını karşılaştırmak ve hesaplar karşısındaki temsilcilere bakmak istiyoruz.

-Önce veri setlerini kontrol etmeliyiz.

-Excel'den bir kişiye tıkla >> Insert >> Pivot table >> Yeni açılan sayfada sağ altta bulunan bölmelere tabloda neyi nasıl görmek istiyorsan o şekilde yerleştir >> excel üzerinde değişiklik yaparsan pivot table sayfasından DATA sonra da refresh all yapıp tabloyu güncelle.

CHECK FOR DUPLICATES

*Data >> Remove Duplicates

*Çoklu değerlerde dublikeleri kaldırmak karmaşık olabilir. Bunun için yeni bir sütun oluşturup fonksiyon girmemiz gerekir:

=if(and(B2kolonu=B3kolonu, C2kolonu=C3kolonu),"Kontrol Et", "")

-Yukarıdaki fonksiyon alt alta 2 veri birbirine eşitse kontrol et olarak bildir; değilse boşluk bırak anlamına geliyor.

*Duplike değerleri formülle ayıklayalım.

-Tüm Kolonu seç >> Conditional Formatting >> Hightlight Cell Rules >> Duplicate values

-Bu bulma işlemi tek kolon veya tek satırda gerçekleşir. Ayrı kolonlarda yer alan isim ve soyisim benzerliklerinde bunu kullanamayız bu yüzden yan kolonda bir değişiklik yapıp tekrar bakmalıyız. Bunun için şu fonksiyonu kullan:

=B2kolonu&C2kolonu >> enter >> şimdi duplikeliğe bakabiliriz. Tüm kırmızıyla işaretlenenleri sildikten sonra tekli isimler kalacak.

*4 farklı kolon varsa ve dublikeleri kaldırmak istiyorsak kendimize kolaylık olması açısından yeni bir duplike kontrol kolonu açmalı ve aşağıdaki örnek formülizasyonu yazmalıyız. Bu bize minik bir kod oluşturacak ve bu kod üzerinden ilerleyeceğiz.

=LEFT(B2,3)&LEFT(C2,2)&MID(D2,2,4)&RIGHT(E2,3)

-B2 sütununda bulunan kelimenin soldan 3 harfini al

-C2 sütununda bulunan kelimenin soldan 2 harfini al

- D2 sütununda bulunan kelimenin ortadan 2,3 ve 4. harfini al
- E2 sütununda bulunan kelimenin sağdan 3 harfini al

2-SPLITTING COLUMNS for DATA CLEAN-UP

**Using Flash Fill:

*PDF'ten veya web üzerinden kopyalanıp excele yapıştırılan veriler genelde tek bir sütunda birleşir. Kendi sıralarında gelirler ama hangi verinin neye ait olduğu karışıktır.

Yan taraftaki boşluğa birkaç veriyi yazınca alta da gölgeli olarak devamı gelir. Enter'a basıp tamamını tek tuşla halledebiliriz.

*Flash fill özelliği veriler tutarlıysa çok yararlı olur fakat tutarsız verileriniz varsa örneğin sadece soy isimleri almak istiyorsunuz ama kişinin 2 ismi var. Bu durumda flash fill özelliği hem 2.ismi hem de soyismi alacaktır dikkat.

**Dynamic Array Functions:

PDF web gibi yerlerden kopyalanıp excele yapıştırılan veriler genelde tek bir sütunda birleşiyordu. Bunları ayırmak için çeşitli fonksiyonlar mevcut:

**Textsplit: Sütunlara ayırır.

=TEXTSPLIT(B2KOLONU," | ",FALSE) enter

-" | " >> bu ifade textte bu şekilde boşluk olarak kullanılmıştı o yüzden böyle yazıldı.

-FALSE >> textteki boş hücreler alınsın

-TRUE >> textteki boş hücreler alınmasın.

**Textbefore:

=TEXTBEFORE(B2KOLONU," ",4 >> ENTER

-" " >> Verileri virgül ile ayır.

-4 >> baştan ilk 4 veriyi al diğerlerini alma.

-Bu ifadeler tekrardan tek bir sütunda toplanır. Yine Textsplit kullanmak gerekir.

=TEXTSPLIT(TEXTBEFORE(B2KOLONU,"",4),",")

****Power Query:**

Sıralaması olmayan, karışık 2den fazla sütunu olan veriler için ayrıştırmada kullanabiliriz.

-Satıra tıklayıp Get Data >> ilgili kolonun tümünü seç >> Split column tıkla >> By Delimiter(sınırlayıcı) >> split by comma(virgül) >>

****Splitting columns into rows****