GOOGLE PLAY APPSTORE

Google play store verilerinin 2018’e kadar tutulduğu bir datasetimiz var şu an. Buradaki amacım “EDA” ve “Feature Engineering” yaparak bu veriyi okumak düzenlemek.

Gerekli kütüphanlerimi dahil ediyorum . Sonra dataseti incelemeye başlıyorum kaç boyutlu olduğuna bakıyorum.

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Daha sonra “info” komudumu çalıştırıp bakıyorum. Bazı datalarımız eksik bazı kolonlarda. “Dtype”na bakınca sadece “Raiting” numeric .

metin, yazılım, bilgisayar simgesi, ekran görüntüsü içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Df.describe() çalıştırdığımızda elimizdeki dataset’e göre sadece “Rating” kolonu numeric. Datamızı daha iyi anlamak için birkaç şeye daha bakmak daha mantıklı oluyor.

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Datamızı incelediğimizde “Reviews” kolonu da numeric olabilcek durumda ilk etapta inceliyorum.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Değerler numeric’e çok yatkın bunun için ilk etapta “dtype”nı değiştiriyorum.

metin, yazı tipi, ekran görüntüsü içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Normalde bu şekilde direkt değişiklik olabiliyor ama 

Bize içerisinde “3.0M” adında numeric’e uygun olmayan bir data olduğunu belirtiyor.

metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Bu komut ile içerisinde “numeric” olmaya meilli “str” değerleri görebiliyoruz

metin, ekran görüntüsü, yazı tipi, makbuz içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Bu şekilde karşımıza “10840” tane numeric olamaya yatkın değer olduğunu belirtiyor ama bizim “10841” tane “Reviews” değerimiz vardı demek ki bir tane değer sadece sorun çıkarmakta. Onu anlamak için ise şunu yapıyoruz;

metin, yazı tipi, çizgi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Bu şekilde “option ü” diyerek ters isareti yapıp “numeric” olmaya ayatkın olmayan verimizi görüyoruz ve tek satır olduğu için onu siliyorum dataset’ten.

metin, ekran görüntüsü, yazılım, sayı, numara içeren bir resim

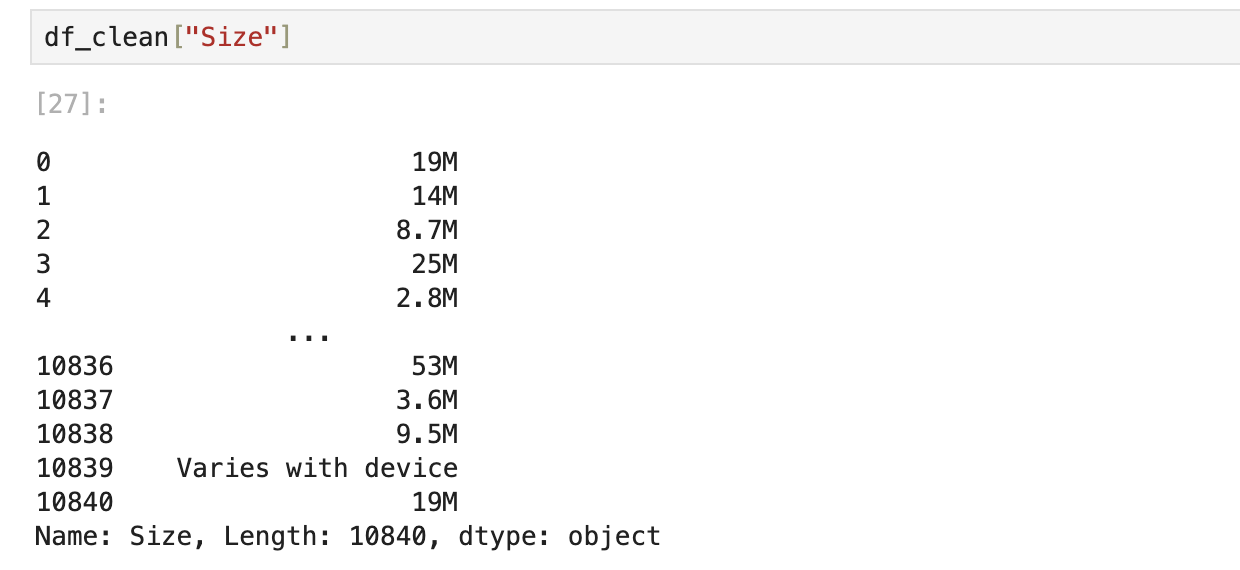
Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Bu şekilde direkt “index” numarası vererek kaldırıyorum.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Bu şekilde “int64” e çevirmiş oldum sonra onun da “describe”nı aldım.

Daha sonra “Size”’mıza bakıyoruz orada da Numeric olabilcek değerler var.  


Önce değerlerime bakıyorum;

metin, ekran görüntüsü, yazı tipi, beyaz içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Bu şekilde değerlerim var . Burada “M” olanlar megabyte zaten onları “k” yapıp sonra “k” olanların hepsini sil diyebilirim.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

3 sıfır eklememin sebebi megabyte, kilobyte’ın 1000 katı .

Daha sonra “k” harflerini de siliyorum.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Bu şekilde “k” harfinden de kurtulduk son olarak “Varies with device” değerleri kaldı onlarda zaten bize net bir değer vermemekte sadece cihaza göre değişir diye bir yazı var bunun yerine “nan” değerlerini girmemiz daha mantıklı olur direkt silersek o kadar row da bosuna gideceği için anlamı kalmaz “(10840 yerine daha düşük)”.

metin, ekran görüntüsü, yazılım, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Daha sonra da numeric hale getiriyorum.  
metin, yazı tipi, sayı, numara, ekran görüntüsü içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Şimdi ise “INTALLS” kolonu ve “Price” kolonu içerisine bakacağım;

metin, ekran görüntüsü, sayı, numara, yazılım içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Şimdi bu şekilde ikisini de incelediğimde çok zor görünmüyor. “Installs” kısmında “+” ve “,” ‘ler silinecek “Price” kısmında da “dolar” işareti silinecek, Bunu tek tek “Size” ‘da yaptığım gibi yapabilirim ama bazen çok fazla işaret ve kolon oluyor onun için direkt “for” loop içerisine alacağım.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Bu şekil ayarlıyorum ve istediğim şekle geliyor.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

En sonda bu şekilde oluyor ve istediğim ayarı yapmış oluyorum.

LAST DATE  
metin, ekran görüntüsü, yazılım, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Şimdi bizim datamızda “last updated” kolunu da object olarak tanımlanmış .  


Biz bunu çok kolay bir şekilde “datetime”’a çevirebiliriz.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Direkt bu şekilde çok kolay şekilde ayarlıyoruz “Last Updated” değerimizi.

Bazen “gün,ay,yıl” bir dataset için ayrı olması önemli olabiliyor bizim su anki datasetimiz için önemli değil ama onları da tek tek bu şekilde yapıyoruz.



Ve bu şekilde çok kolay bir biçimde hallediliyor.

Daha sonr DUPLICATED olanlara bakıyoruz bazı veriler birden fazla kez yazılmış olabilir.

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Ve bu şekilde 1181 tane “duplicate” verimiz çıkıyor.  


Bu şekilde siliyoruz “dublicate” olanları.

Daha sonra “numeric” olan verilerin tablolarını incelemek isityorum bunun için önce numeric ve categoric değerleri ayırıyorum.



Bu şekilde [feature for feature] yazmasının mantıgı;  
liste = []

for i in [1, 2, 3]:

liste.append(i)

Bunla aynı direkt “features” diye bir listeye atmıs oluyoruz. metin, ekran görüntüsü, makbuz içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Daha sonra bu şekilde oluyor.

diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi, ekran görüntüsü içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Daha sonra bu şekilde grafiklerimi oluşturuyorum.

ekran görüntüsü, metin, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Bu şekil belli başlı kategorileri oluşturuyorum . şimdi hangi ürün kategoriisde en çok satış var bunu bulmaya çalışacağım.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Bu şekilde yapınca en çok hangi ürün kategorisi var bulmuş oluyor hangi kategori daha fazla indirilmiş anlamıyoruz.  
metin, yazılım, web sayfası, web sitesi içeren bir resim

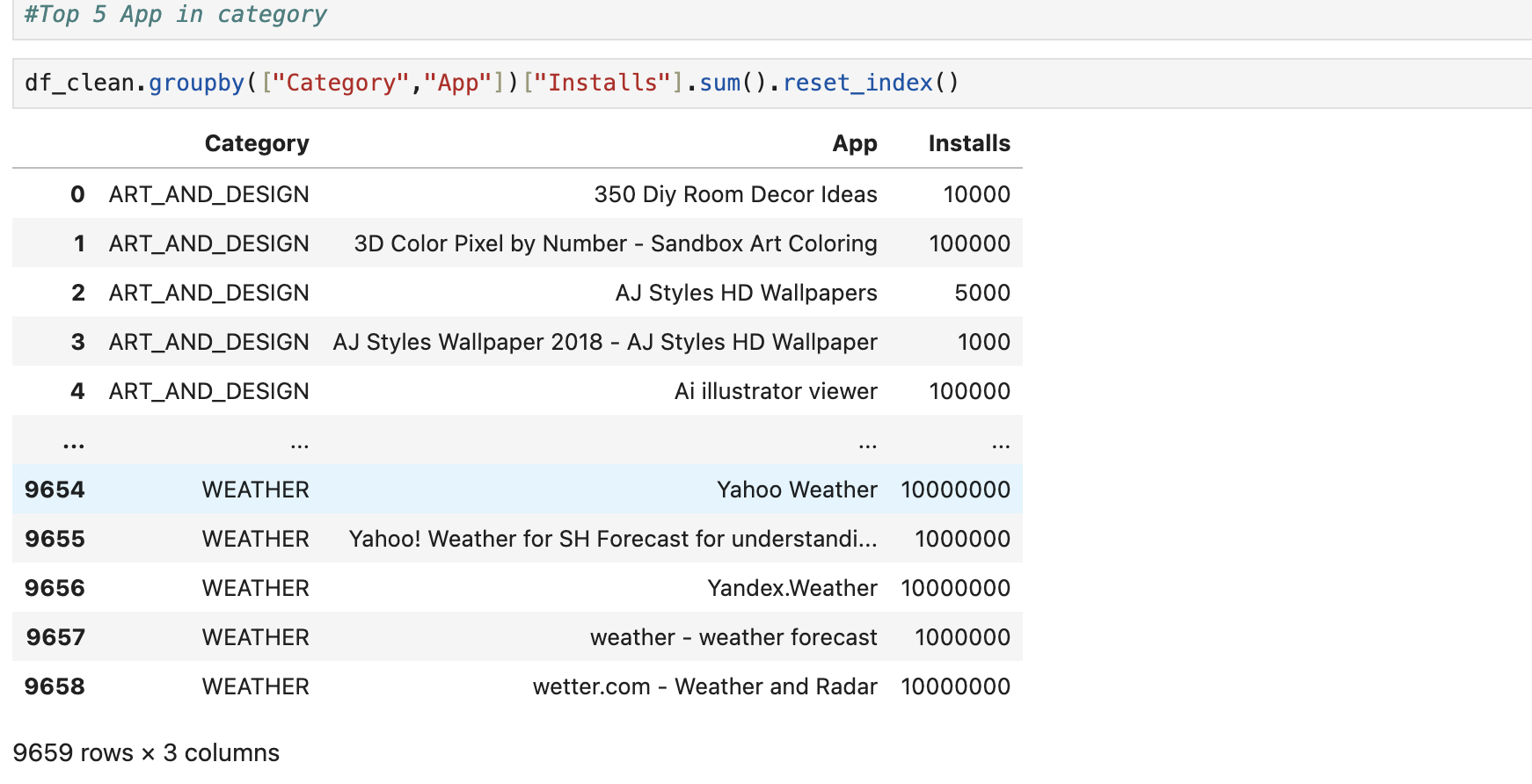
Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Bu şekilde “group by” kullanarak ayarlıyorum.

metin, ekran görüntüsü, yazılım, web sayfası içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Daha sonra bu şekilde grafik halini çıkarıyorum.

Şimdi en cok indirilen kategorilerdeki ilk 5 app’İ getirmeye çalışacağım.  


İlk etapta bu şekilde “Category ve App”’e göre gruplayıp “Installs”’ını alıyorum. “reset\_index” dememin sebebi düzgün bir df olması için yoksa “Series” olarak kalır.

Daha sonra;  
metin, yazı tipi, sayı, numara, ekran görüntüsü içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

“Installs”’a göre büyükten küçüğe sıralıyorum. Benim amacım ilk 5 kategorideki kategorilere göre ilk 5 app’i almak olduğu için ilk 5 kategoriyi bir diziye atıyorum.

metin, ekran görüntüsü, çizgi, diyagram içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Bu şekil for loop içerisine alıyorum. Enerumete’ın normal “for’dan” hiçbir farkı yok sadece indexleri belirtmek istiyoruz index’i belirtmek için de ekstra bir parametre eklemiş oluyoruz mesela burada “i”’i ekledik. Bu şekilde grafiğimiz oluşuyor.