Popular Video Games – Data Science with Python Final Projesi Raporu – Elif Beyza Gökdemir

**Proje Konusu :** Bu projede konu seçerken hem keyif aldığım hem de üzerinde biraz bilgim olduğu popüler oyunlar konusunu seçmek istedim. Steam, Epic Games gibi birkaç platformdan arada oyunları gözlemleme imkanım olduğu için ve de kendim de çeşitli oyunlar denemeyi sevdiğim için bu konuda uygun dataset aramaya başladım. Aradığım datasetlerde aslında genel bulmak istediğim veriler öncelik olarak ‘kullanıcıların yaş ortalaması’ , hangi bölgede ne kadar oynanmış gibi bilgiler idi ancak hiçbirinde kullanıcı profillerine dair bilgiler yoktu, aklıma gelen başka proje fikirlere göre datasetlere baktığımda ise hiç uygun veriler bulamadım, bu projede kullandığım dataset diğerlerine nazaran çok daha verimli duruyordu o yüzden son kararım bu dataset oldu.

**Data Temizleme :** Dataset’in discussion kısmını incelediğimde Times Listed ve Number of Reviews sütunlarının tamamen aynı olduğunu gördüm, bu yüzden bir tanesini çıkarmak zorunda kaldım. Unnamed: 0 isimli index ile tamamen aynı bir sütun daha vardı, tamamen gereksiz olduğu için çıkarttım. Son olarak , Summary ve Review ı çıkarttım çünkü bu projemde bana katkı olmayacağını düşündüm, bu sütunların içinden belli kelimeler aratarak başka sonuçlar da elde edilebilirdi ancak bu proje için fazla detaylı ve zaman alıcı olacağını düşündüğüm için bu verileri tamamen çıkartmaya karar verdim.

Sonrasında hangi verilerde NaN değerlerin bulunduğuna baktım. Sadece ‘Rating’ ve ‘Team’ sütunlarında vardı.

Team : sadece bir oyunda NaN değer vardı, team önemli bir veri olduğu için ve bu oyun oldukça az sayıda oynanmış olduğu için çıkardım.

Rating : Datasette henüz çıkmamış ya da bu dataset oluşturulduğunda daha yeni çıkmış oyunlar vardı, o yüzden rating değerleri NaN idi o yüzden bu oyunları da çıkardım.

Ayriyeten Release Date’leri 'releases on TBD' olarak gözüken iki oyun daha vardı. Bunları incelediğimde de bir tanesinin daha çıkmamış olduğunu, öbürünün ise verisinin eksik olduğunu fark ettim. Çıkmamış olanı çıkarttım, verisi eksik olanın ise çıkış tarihini bulup veriye ekledim.

Review sütunundan dolayı bazı oyunlar birden fazla review aldığı için duplicate olarak gözüküyordu, o yüzden bütün duplicate değerleri de temizledim.

Sayısal değerler bin üzerinden gösterilmişti (mesela 3000 = 3K) ve object tipinde gözüküyorlardı, numerik olan sütunları numerik değere dönüştürdüm. Aynı şekilde oyunların yayınlanma tarihlerinden faydalanabilmek için Release Date’i de dönüştürdüm.

Datasette oldukça fazla aykırı değer olduğu için bunları filtreleyerek birkaç ayrı versiyon dataframeler daha oluşturdum. Bazı sorularımda bu dataframeleri kullandım, bazılarında da karşılaştırmak istedim.

**Sorular:**

**1) Hangi oyuna en fazla review yapılmış? En fazla review yapılan oyun nasıl bir oyun , ne kadar seviliyor ve ratingi ne kadar yüksek görmek istedim.**

**2) Oyunların Rating dağılımı ve grafiği nedir? (%10 quantile dahil ederek ve etmeyerek) Oyunlara genel olarak baktığımızda nasıl puanlar alıyorlar bunu toplam oynanma sayısının %10 quantileden aşağısını da dahil ederek ve etmeyerek bakmak istedim.**

**3) 4'ün üstünde oylama alan kaç oyun vardır?**

**4) En çok güncel olarak hangi oyun oynanmaktadır? Şu an en çok oynanıyor olan oyunun ne kadar eski olup puanı kaç olduğu gibi verilerin katkısı olacağını düşündüm.**

**5) Oyunların yayınlanma yılına göre ortalama Rating ve Plays dağılımı nedir? Oyunların ortalama oynanma sayısına yıl olarak baktığımda 1985’in bariz bir şekilde ortaya çıktığını çünkü o yıl çıkan sadece tek bir oyun olduğunu fark ettim. Bu yüzden bir sonraki sorum belli olmuş oldu.**

**6) Oyunların çıkış yılına göre toplam ne kadar oynanmaya sahip olduğunun ve hangi yıl ne kadar oyun çıktığının grafikleri nedir? Burada günümüze yaklaştıkça çıkan oyun sayısının oldukça arttığını gözlemledim.**

**7) Hangi oyun en çok backlogs/plays oranına sahiptir? Bir oyunun toplam oynanma sayısına nazaran backlog sayısı daha fazlaysa oyun için muhtemelen koleksiyon için indirimde alınmış ya da ücretsiz bir oyun olabilir. Hangi oyun için bu oran en fazla gözlemlemek istedim.**

**8) Güncel olarak ortalama ne kadar oyun oynanmaktadır ve oynanmıştır?**

**9) En yüksek oynanmaya sahip oyunun ratingi nedir?**

**10) Nintendo tarafından yapılmış oyunların rating ortalaması Datasette en fazla oyun yayınlamış olan Nintendo şirketinin oyunlarının rating ortalamasını aldığımızda, aslında genel olarak ne kadar beğeniliyor ve bu kadar oyun çıkarmış olmanın puanlama üzerinde nasıl bir etkisi oluyor incelemek istedim.**

**11) En yüksek ratinge sahip oyunlar nelerdir?**

**12) Oynanma sayısı ile backlog sayısı kıyaslama**

**13) En çok oynanan 10 oyunun wishlist grafiği Bu grafik de aslında bize kısmen bu oyunların ücreti ve boyutu hakkında bilgi verebiliyor.**

**14) Hangi ayda ne kadar oyun çıkmış? Genelde oyunların yüksek oranda Eylül-Kasım aralığında çıktığını görüyoruz. Sonbahar mevsiminin oyun geliştiriciler için önemli bir zaman olduğunu söyleyebiliriz.**

**15) En çok tercih edilen Janralar ve grafiği**

**16) Veri kümesinde kaç oyun şirketi bulunmaktadır ve kaç oyunları vardır?**

**17) En çok hangi takımın oyunları wishliste alınmıştır?**

**18) Ortalamadan fazla review sayısı olan oyunların oyun türü dağılımı nasıldır?**

**19) Playing/Plays oranına göre en yüksek orana sahip ilk 10 oyun nedir? Hangi oyunlar en çok ‘sadık’ oyuncuya sahip? Bu konuda bilgi alabilmek adına bu oranı incelemek istedim.**

**20) En çok backloga alınmış 50 oyunun Rating ortalamaları nedir? Bu oyunlar koleksiyonluk mu alınıyor yoksa daha sonra oynayabilmek için indirimdeyken mi yakalanıyor yoksa ikisi de mi? Bu listeye baktığımızda birçok sevilen oyunun listede yer aldığını görüyoruz.**

**Machine Learning : Bu kısımda tahmin olarak Ratingi kullanmanın en kullanışlı olacağını düşündüğümden Rating seçtim. Değerimiz continuous olduğu için lineer regresyon kullandım. Veriyi kullanmaya başlamadan önce, Release Date’den faydalanarak oyunların gün bazında yaşını çıkarttım ve 2023te çıkan oyunları dahil etmedim, Genre kısmından sadece Indie mi değil mi diye bakmaya karar verdim ve Team sütununu katmadım, aksi halde dataframe oldukça büyüyecek ve çok daha fazla aksiyon gerektirecekti, zamanımın yetmeyeceğini düşündüğümden dolayı olabildiğince verimli bir şekilde filtrelemeye çalıştım. Çok verimli bir training olmasa da 0.3 MAE ile elimizdeki verilerin kısıtlılığına bakarsak yine de fena olmayan bir sonuç çıkarttığını düşünüyorum.**