**Netflix Originals Veri Setinin Analizi**

**Ekip üyeleri**: Gaye Çetindere, Ecem Çırakoğlu, Talha Enes Baba ve Furkan Temel

**Projenin amacı**: Netflix Originals veri setini görselleştirme teknikleriyle birlikte analiz etmektir.

**Kullanılan Programlama Dili:** Python

**Proje Tanımı:**

GlobalAiHub python bootcamp sürecinin bir projesi olarak Netflix Originals veri setinin analizine ilişkin istenilen maddeler aşağıdaki gibidir.

* Veri setine göre uzun soluklu filmler hangi dilde oluşturulmuştur? Görselleştirme yapınız.
* 2019 Ocak ile 2020 Haziran tarihleri arasında 'Documentary' türünde çekilmiş filmlerin IMDB değerlerini bulup görselleştiriniz.
* İngilizce çekilen filmler içerisinde hangi tür en yüksek IMDB puanına sahiptir?
* 'Hindi' Dilinde çekilmiş olan filmlerin ortalama 'runtime' suresi nedir?
* 'Genre' Sütunu kaç kategoriye sahiptir ve bu kategoriler nelerdir? Görselleştirerek ifade ediniz.
* Veri setinde bulunan filmlerde en çok kullanılan 3 dili bulunuz.
* IMDB puanı en yüksek olan ilk 10 film hangileridir?
* IMDB puanı ile 'Runtime' arasında nasıl bir korelasyon vardır? İnceleyip görselleştiriniz.
* IMDB Puanı en yüksek olan ilk 10 'Genre' hangileridir? Görselleştiriniz.
* 'Runtime' değeri en yüksek olan ilk 10 film hangileridir? Görselleştiriniz.
* Hangi yılda en fazla film yayımlanmıştır? Görselleştiriniz.
* Hangi dilde yayımlanan filmler en düşük ortalama IMBD puanına sahiptir? Görselleştiriniz.
* Hangi yılın toplam "runtime" süresi en fazladır?
* Her bir dilin en fazla kullanıldığı "Genre" nedir?
* Veri setinde outlier veri var mıdır? Açıklayınız.

Projede maddelerin yukarıdaki yazılış sıraları dikkate alınmıştır. Pandas, Numpy, Matplotlib ve Seaborn kütüphanelerinden yararlanılmıştır. Netflix Originals veri setinin ismi “*df*” olarak atanmıştır.

**1. madde:**

Uzun soluklu filmleri, kısa soluklu filmlerden ayırt etmek için Sinema Sanatları ve Bilimleri Akademisi, Amerikan Film Enstitüsü ve Britanya Film Enstitüsü’nün uzun metraj film tanımları dikkate alınmıştır. Söz konusu tanımlara göre 40 dakikanın üstündeki filmler uzun metraj film kategorisine girmektedir. Görselleştirme için çubuk grafiği kullanılmıştır.

**2. madde:**

Veri setindeki filmlerin çıkış tarihini gösteren “*Premiere*” sütununun zaman biçimi değiştirilerek “*pandas*” kütüphanesinde analize uygun, “*date*” isimli yeni bir zaman sütunu veri setine eklenmiştir. “*date*” sütunundan yararlanarak, 2019 Ocak ve 2020 Haziran tarihleri arasında “*Documentary*” türündeki filmlerin IMDB puanlarının içeren “*doc\_btw\_2019\_2020*” adında yeni bir veri seti oluşturulmuştur. Görselleştirme için çubuk grafiği kullanılmıştır.

**3. madde:**

“*df\_new*” adında İngilizce dilinde çekilen filmlerin türlerine göre sınıflandırıldığı ve söz konusu türün en yüksek IMDB puanının yer aldığı yeni bir veri seti oluşturulmuştur.

**4. madde:**

“*time*” değişkeni, “*Hindi*” dilinde çekilmiş filmlerin ortalama “*Runtime*” süresini ifade etmektedir.

**5. madde:**

Söz konusu maddede istenilen bilgiye ulaşmak için “*count\_plot*” fonksiyonu oluşturulmuştur.

Fonksiyonda tanımlanan “*how\_many*” değişkeni, “*df*” veri setinden seçilen sütunun özgün değerlerinin sayısını vermektedir. Bu madde de “*Genre*” sütununda kaç tane kategori olduğu bilgisine ulaşmak için kullanılmıştır.

“*cat*” değişkeni ise “*df*” veri setinden seçilen sütuna ait özgün değerlerin kaç kez tekrar ettiği bilgisini vermektedir. Söz konusu madde için “*Genre*” sütunundaki her bir film kategorisinin ne kadar tekrar ettiği bilgisini verir.

Görselleştirme için ise “*sns.countplot*” fonksiyonu ile “*seaborn*” kütüphanesinden yararlanılmıştır.

**6. madde:**

“*df*” veri setinde birden fazla dil içeren filmlerin dillerinin ayrıldığı ve her bir filmin tek bir dile sahip olduğu “*df\_l*” adında yeni bir veri seti oluşturulmuştur.

**7. madde:**

“*pandas*” kütüphanesindeki “*DataFrame.nlargest()*” fonksiyonundan yararlanılmıştır.

**8. madde:**

“*df*” veri setine ilişkin korelasyon değerlerini ölçmek için “*korelasyon*” fonksiyonu oluşturulmuştur. Fonksiyonun içerisindeki “*num\_col*” değişkeni veri setinden sadece sayısal değişkenlerin dikkate alınmasında rol oynar. “*corr*” değişkeni ise veri setinden sayısal değerler içeren sütunlar korelasyonunu hesaplar. Görselleştirme için ısı haritası kullanılmıştır.

**9. madde:**

“*df*” veri setinin IMDB puanına göre azalan şekilde sıralandığı “*sort\_by\_imdb*” isimli yeni bir veri seti oluşturulmuştur. Daha sonra “*sort\_by\_imdb*” veri setinin “*Genre*” sütunundaki ilk 10 özgün kategoriden oluşan “*top\_ten\_genre*” isimli yeni bir dizi oluşturulmuştur. “*top\_ten\_genre\_index*” listesi ile veri setindeki IMDB puanı en yüksek 10 farklı “*Genre*” kategorisinin eşleştirildiği “*imdb*” veri seti elde edilmiştir. Görselleştirme çubuk grafiği ile gerçekleştirilmiştir.

**10.madde:**

9. maddeyle benzer şekilde bu sefer “*Runtime*” sütununa göre azalan şekilde sıralanan “*sort\_by\_runtime*” isim yeni bir veri seti oluşturulmuştur. Bu yeni veri setinin ilk 10 “*Title*” sütunundaki film isimleri “*highest\_runtime\_names*” olarak, ilk 10 “Runtime” sütunundaki veriler ise “*highest\_runtime*” olarak adlandırılmıştır. Görselleştirme için çubuk grafiğinden yararlanılmıştır.

**11.madde:**

“*df*” veri setinin “*Premiere*” sütunu pandas kütüphanesine uygun hale getirilerek “*year*” veri seti oluşturulmuştur. “*year*” veri setinden yararlanarak görselleştirmek üzere x ve y eksenleri için sırasıyla “*year\_index*” ve “*year\_values*” dizileri oluşturulmuştur. Görselleştirme için çubuk grafiği tercih edilmiştir.

**12.madde:**

“*df*” veri seti dillere göre gruplandırılıp her bil dil için yayımlanan filmlerin ortalama IMDB puanı hesaplanmıştır. X ekseni diller, y ekseni IMDB puanları olacak şekilde çubuk grafiği ile görselleştirilmiştir.

**13.madde:**

“*df*” veri setinin “*Premiere*” sütununun biçimlendirilmiş halini içeren “*splitted\_premiere*” veri seti elde edilmiştir. Bu veri seti “*grouped\_years*” değişkeni ile yıllara göre gruplandırılmış biçimde “*Runtime*” süreleri gözlemlenmiştir. “*max\_runtime*” ile en fazla film süresi ne kadar olduğu ve hangi yıl olduğu tespit edilmiştir.

**14.madde:**

“*df\_new*” değişkeni ile birden fazla dil içeren filmlerin ayrıldığı yeni bir veri seti oluşturulmuştur. “*df\_new2*” değişkeni ile “*df\_new*” veri seti diller ve türlere göre kategorize edilmiş ve istenilen çıktıya ulaşılmıştır.

**15.madde:**

“*num*” değişkeni ile “*df*” veri setinden sayısal verileri içeren sütunları içeren veri seti oluşturulmuştur. “*outlier\_val*” fonksiyonu ile “*Runtime*” ve “*IMDB score*” verilerine göre outlier değerleri hesaplanmıştır.