|  |
| --- |
|  |
| * Volkan Cebeci – [cebecivolkan34@gmail.com](mailto:cebecivolkan34@gmail.com) -https://github.com/VolkanCebeci - <https://www.linkedin.com/in/volkan-cebeci-23a863223/> * İsmail Sağdıç-isagdic@yahoo.comhttps://github.com/ismailsagdic * Merve Altın-altinmerve@outlook.com.tr-https://www.linkedin.com/in/mervealtin1/ - https://github.com/mervealtin * Ecem Birsöz-ecemmbirsoz@gmail.com |





Netflix Original Films & IMDB Scores Data Analysıs Report

# Global AI Hub Python Bootcamp

## NETFLİX ORİGİNAL FİLMLERİ & IMDB SKORLARI VERİ ANALİZİ RAPORU

### **Giriş**

Bu çalışmada Netflix Original’in içeriğinde bulunan filmlere ait verilere dayanılarak oluşturulan analizle çeşitli değişkenlere etkilerini gözlemleyip istatistiksel değerlerini inceleyeceğiz. Çeşitli grafik ve veri analizi metotlarıyla daha okunabilir hale getirip görsellik kazandıracağız. Bu analizde python programlama dili kullanılarak veri tabloları oluşturulmuştur. Kategorik ve sayısal değişkenler arasındaki ilişkiler gözlemlenmiş, yorumlanmış ve görselleştirilmiştir.

Kütüphanelerin İçe Aktarımı

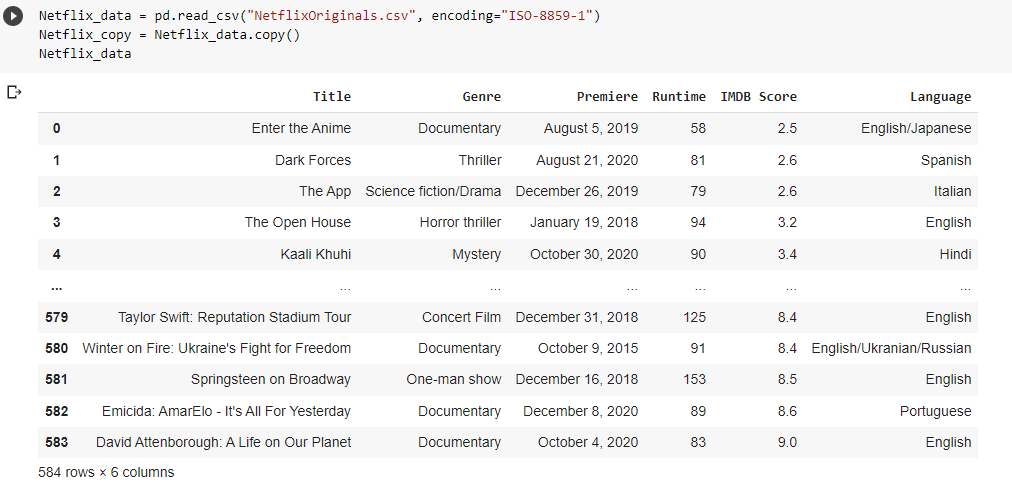
metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

“Görsel 1.0 Python kütüphanelerin yüklenmesi”

Veri Ön İnceleme

tablo içeren bir resim

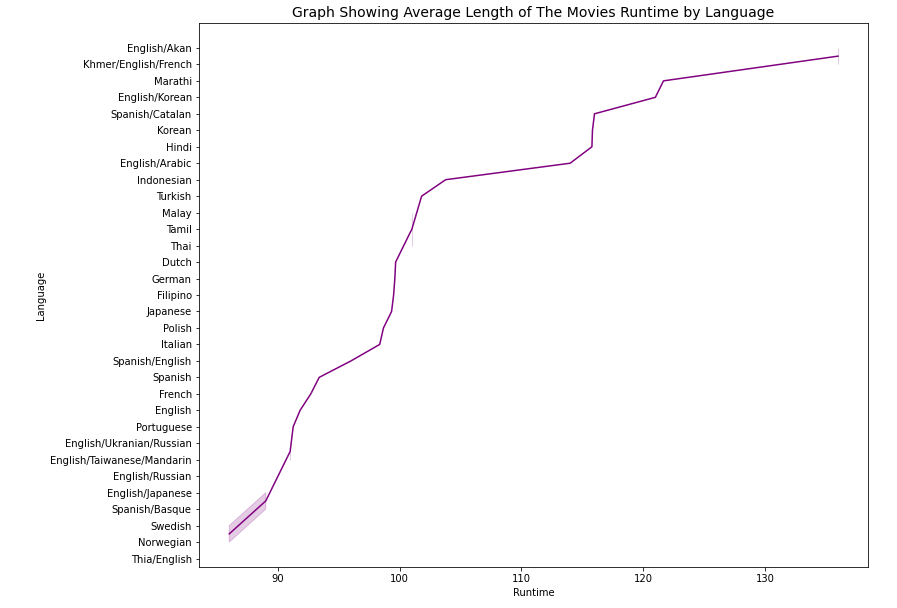
Açıklama otomatik olarak oluşturuldu

“Tablo 1.0 Netflix Original veri seti” “Görsel 1.1 veri setine yönelik bilgiler”

* Veri tablosunda oluşan değerlere göre 27.5+ Kib boyutunda bir veri olduğu ve veriler arasında eksik değer olmadığı görülmektedir.

Veri Analizi

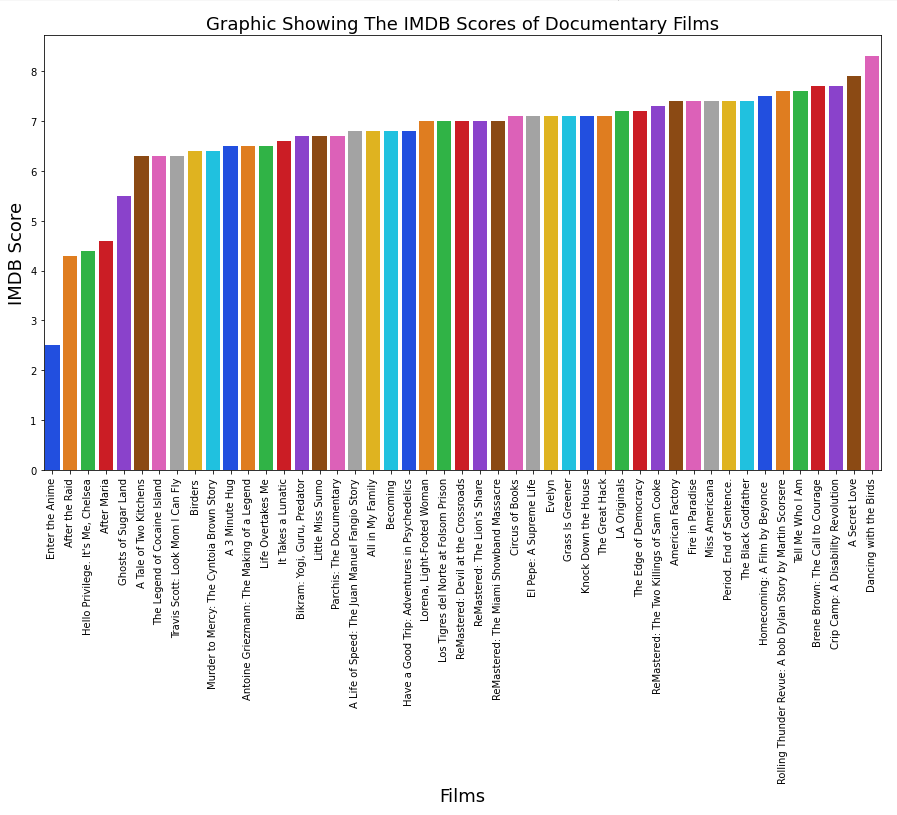
* Veri setine göre uzun soluklu filmlerin hangi dilde oluştuğuna dair listeleyelim ve görselleştirelim.
* Uzun soluklu çok açık bir yorum olduğundan sezgisel olarak 86 dk’dan fazla süren filmleri listeleyip görselleştirelim.
* Grafikte “English” dilindeki filmlerin daha uzun soluklu olduğunu görebiliriz.

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

“Görsel 1.2 Uzun soluklu filmlerin ortalama süreleri” “Görsel 1.3 Ortalama Runtime-Language çizgi grafiği

* 2019 Ocak ile 2020 Haziran tarihleri arasında 'Documentary' türünde çekilmiş filmlerin IMDB değerlerini bulup görselleştirelim.



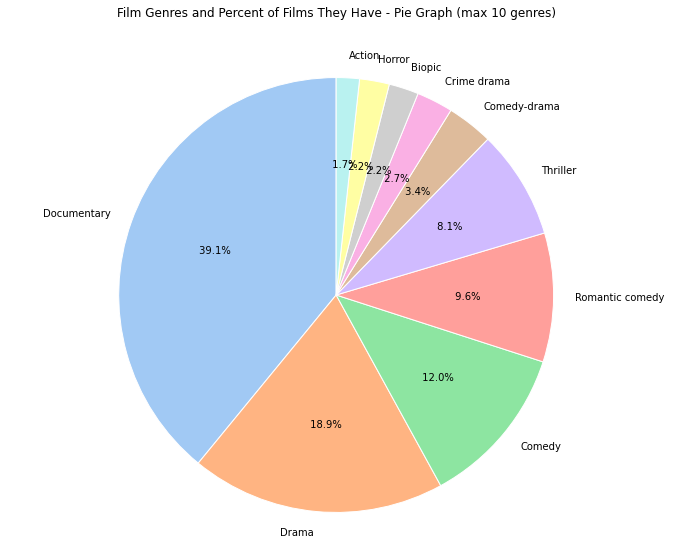
“Görsel 1.4 2019 Ocak ile 2020 Haziran Film-IMDB sütun grafiği”

* İlgili tarihler arasındaki “Dancing with the Birds” filmi 8.3 IMDB skoruyla en fazla “Enter the Anime” filmi 2.9 IMDB skoruyla en az puanı elde ettiklerini görüyoruz.



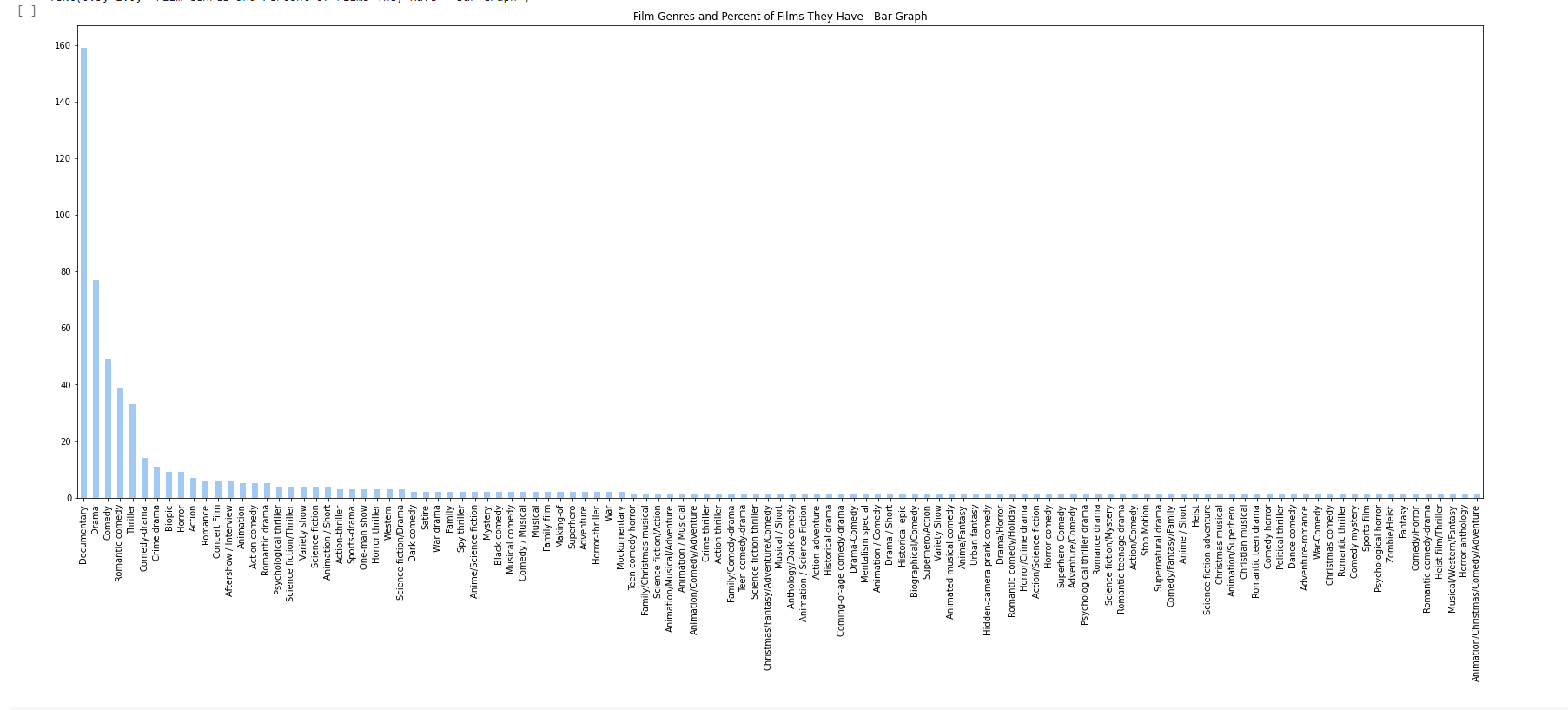
“Tablo 1.1 İngilizce dilindeki filmlerin en yüksek IMDB skoru sıralaması”

* Yukarıdaki tabloda İngilizce dilindeki filmlerden en yüksek IMDB skoruna ait sıralamayı tablo halinde görebiliriz.
* Netflix Original türlerine göre toplam 115 farklı yapıma ev sahipliği yapmıştır.



“Görsel 1.5 Filmlerin yüzdelik türlerine göre dağılım grafiği”

* Daire grafiğinde ilk 10 yapımın türüne göre yüzdelik dağılımını incelersek pastanın yarısından fazlasını %39.1 ile belgesel, %18.9 ile drama, %12.0 ile komedi türleri oluşturuyor.



“Görsel 1.6 Türlerine göre filmlerin sütun grafiği”

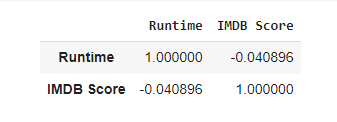
* Sütun grafiğine bakarak daha fazla geniş aralıkta birçok farklı çeşidin bulunduğunu gözlemleyebiliriz.
* Veri setine göre en çok kullanılan diller 401 kez İngilizce, 33 kez Hindice, 31 kez ise İspanyolca dillerinden oluşmuştur.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

“Görsel 1.7 IMDB skoruna göre liste”

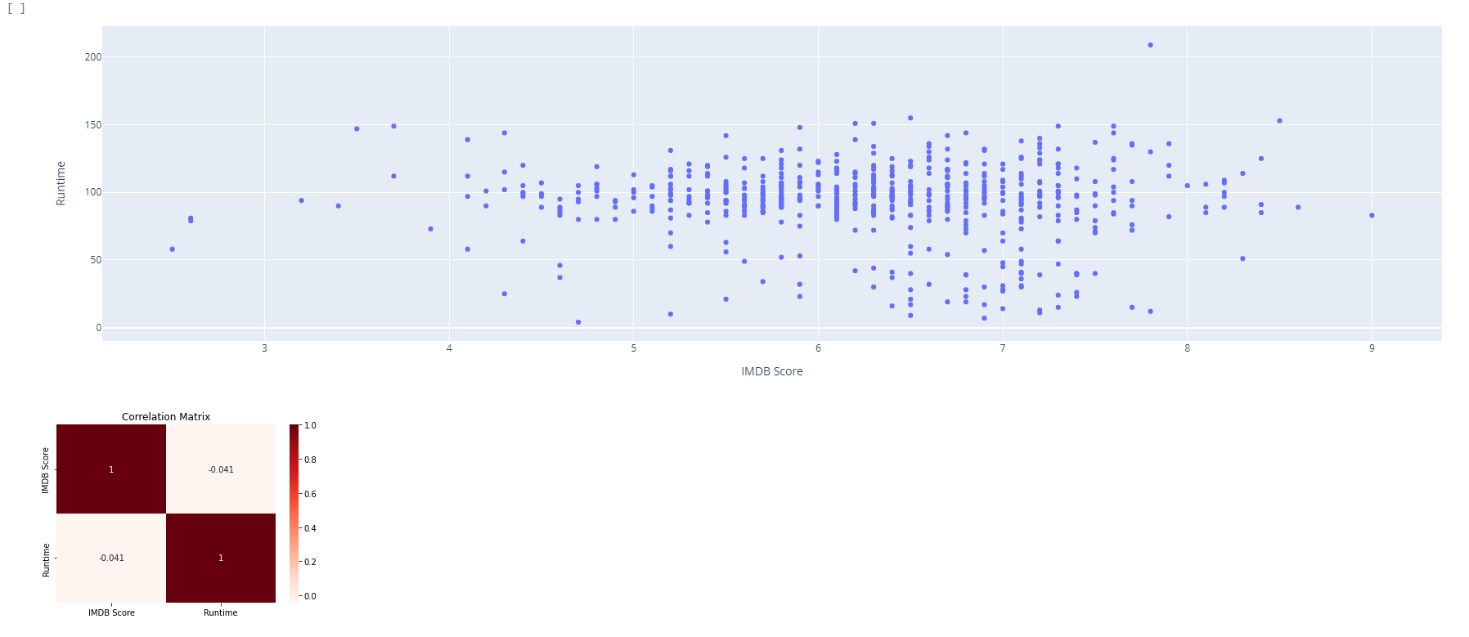
* IMDB skoruna göre ilk 10 film listelenmiştir.



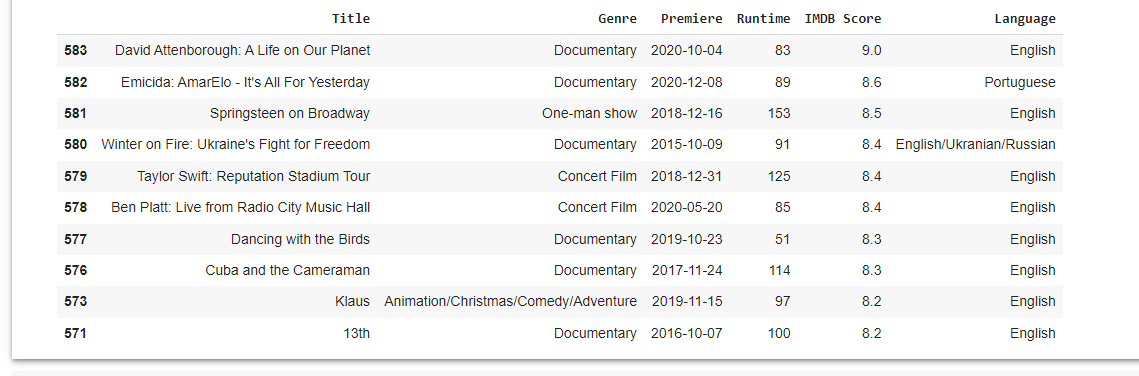
“Tablo 1.2 IMDB-Runtime arasındaki korelasyon değerleri”

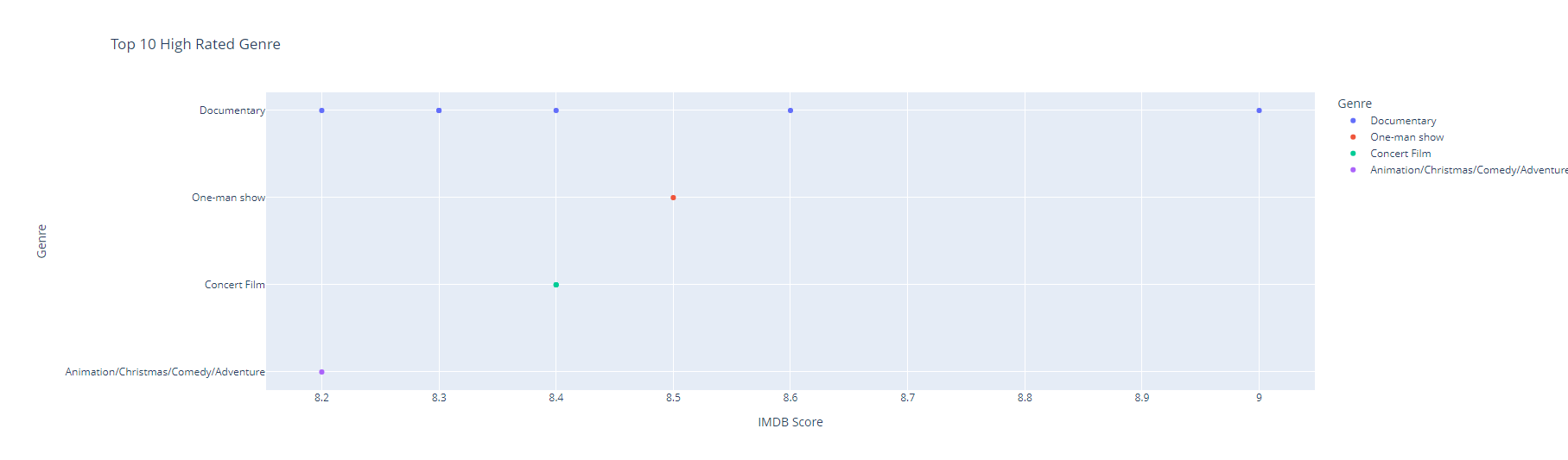
* **Korelasyon katsayısı (r) nın yorumu;**
  + r<0.2 ise çok zayıf ilişki yada korelasyon yok
  + 0.2-0.4 arasında ise zayıf korelasyon
  + 0.4-0.6 arasında ise orta şiddette korelasyon
  + 0.6-0.8 arasında ise yüksek korelasyon
  + 0.8> ise çok yüksek korelasyon olduğu yorumu yapılır.
* Bilgilerden yola çıkıp istenilen durumu yorumlayacak olursak;

“Runtime” ve “IMDB Score” negatif ilişkili gibi görünen çok zayıf bir korelasyon olduğunu veya herhangi bir korelasyon olmadığını söyleyebiliriz. Yani Runtime değerinde olası bir artış “IMDB Score” değerinde bir artış veya bir azalışa sebebiyet vermemekte.



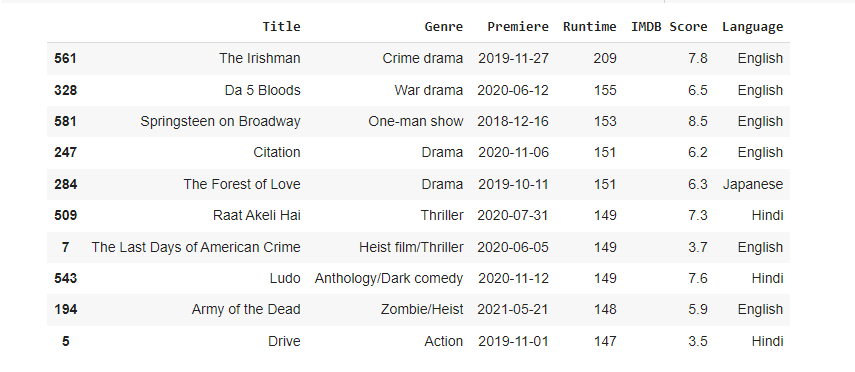
“Görsel 1.8 IMDB-Runtime arasındaki korelasyon dağılım graği ve ısı grafiği”



“Tablo 1.3 IMDB skor listesi”

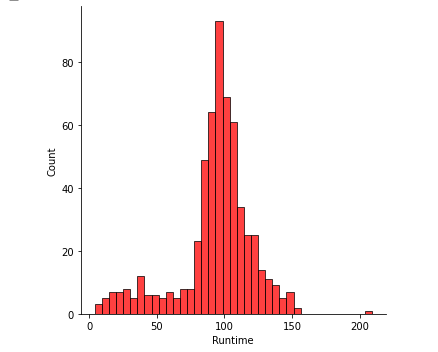
“Görsel 1.9 IMDB skor dağılım nokta grafiği”

* Yukarıdaki grafikte IMDB skoru en yüksek ilk 10 filmi ve bunların türleri arasındaki ilişkiyi görebiliriz.

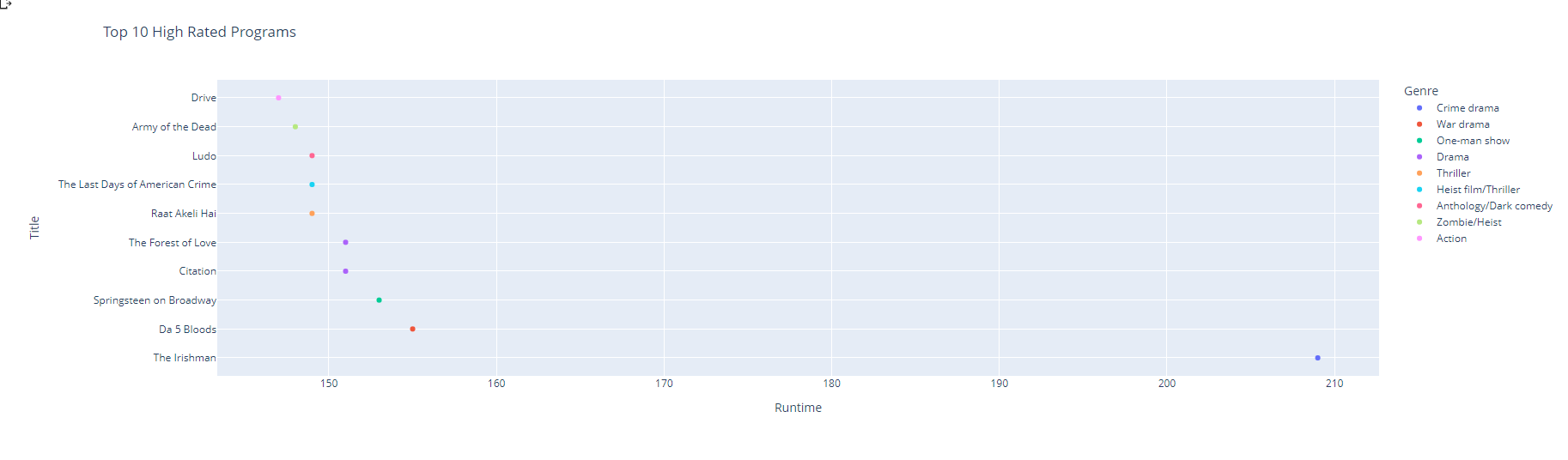


“Tablo 1.4 Runtime uzunluğu listesi”

* Filmlerin gösteri uzunluğuna göre ilk 10 film sıralanmıştır. Sütun grafiğinden de anlaşılması üzerine İngilizce dilindeki filmler daha fazla uzunluğa sahip filmler içeriyor.

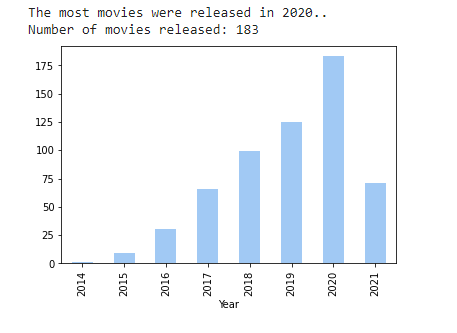


“Görsel 1.10 Runtime sütun grafiği”



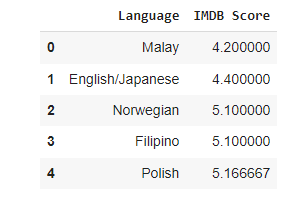
“Görsel 1.11 Runtime dağılım nokta grafiği”

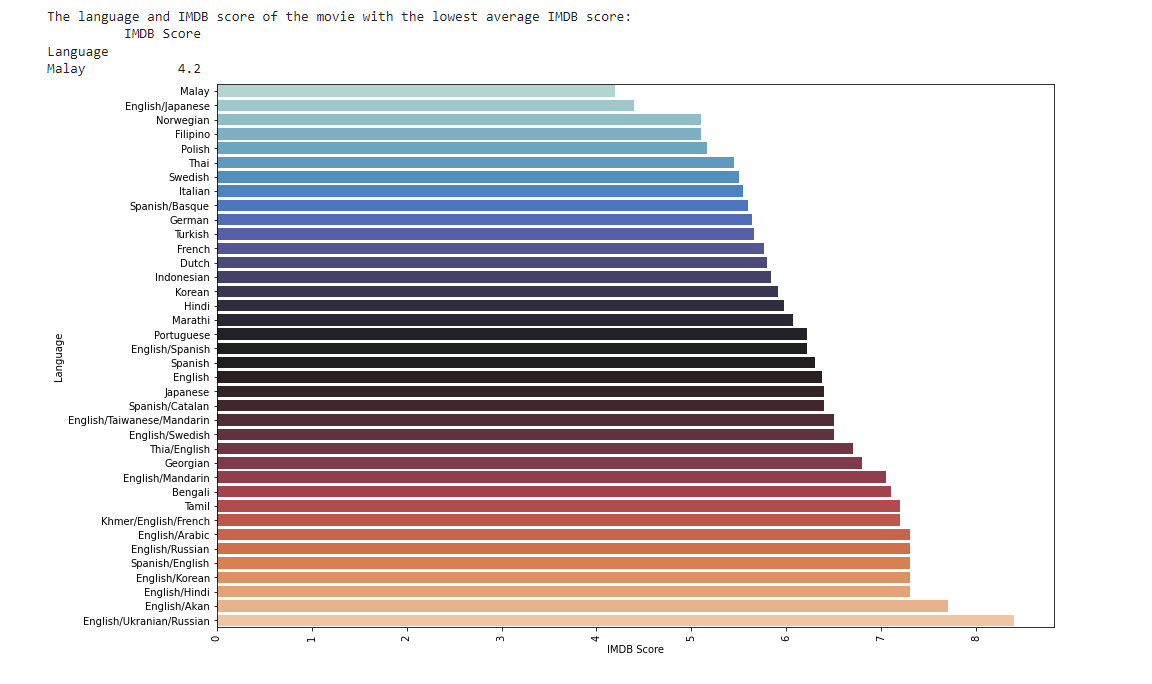
* Nokta grafiğinde Crime&Drama türündeki filmin gösteri uzunluğuna göre daha ayrıldığını görebiliriz.



“Görsel 1.12 Yıllara göre film yapım sütun grafiği”

* Veri setindeki tüm yılları baz aldığımızda en fazla 2020 yılında 183 içerik kullanıcı ile buluşmuştur.



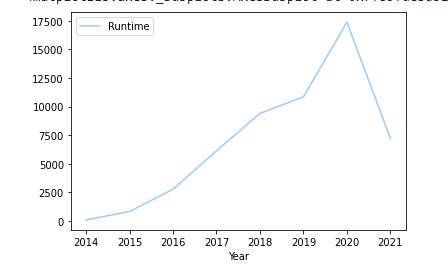
 “Tablo 1.5 En düşük IMDB skora sahip diller”

“Görsel 1.13 IMDB Skor ortalaması-Dil sütun grafiği”

* Yayınlanan filmlerin ortalamalarına göre en düşten en büyüye sıralanmasına yönelik sütun grafiğinde soğuktan sıcağa renklendirilerek ilişkilendirilmiştir. Ortalamaya göre “Malay” dili en düşük “English\Ukrainan\Russian” dilleri en fazla IMDB skoruna sahip olmuşlardır.

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu



“Tablo 1.6 Yılın toplam film süreleri” “Toplam film sürelerine göre çizgi grafiği”

* Tabloya ve çizgi grafiğine göre veri setindeki tüm yılların toplam içeriklerindeki filmlerin gösteri uzunluk süreleri en yüksek 2020 yılı olmak üzere 17384 olarak ölçülmüştür. Sonraki yıl ani bir düşüş gözlemlenmiştir.

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

“Görsel 1.14 Türlerine göre en çok izlenen filmlerin dilleri listesi”

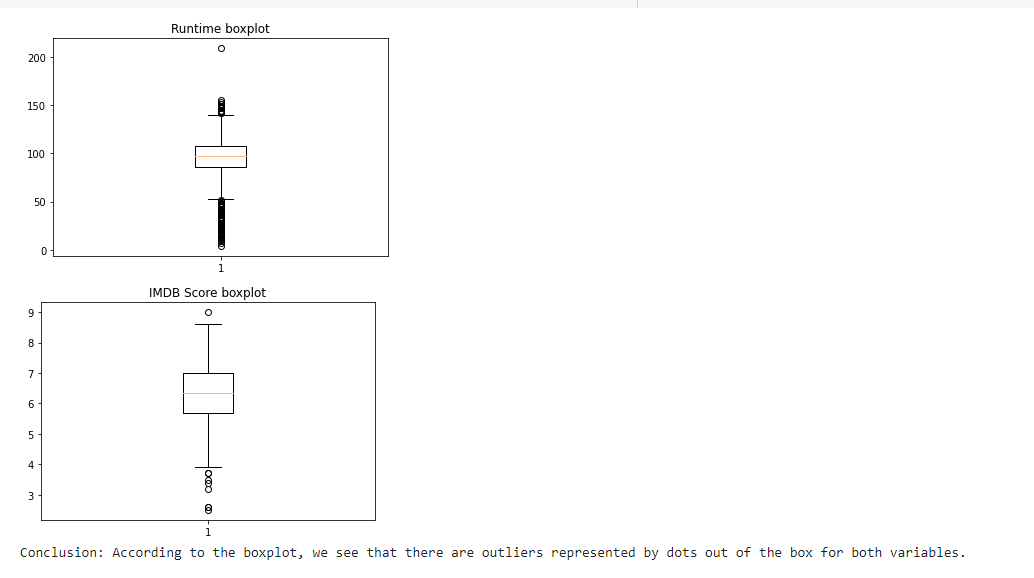
* Türlerine göre her bir dilin en çok kullanıldığı listelenmiştir.

Sonuç

* Son olarak veri setimizde aykırı bir veri olup olmadığını gözlemleyelim.
* Aykırı Gözlem (Outlier): Diğer gözlemlerden kayda değer derecede uzak olan gözleme aykırı veya uç değer denir.
* Aykırı Değerler neden sorun yaratır?
  + Verilerin dağılımını ve ortalama, medyan vs. gibi veriyi temsil eden istatistikleri etkiler.
  + Modellerden elde edilen sonuçlara etki eder.
  + İstatistiksel testlerin gücünü düşürür.
* Değişken sayısının az olduğu durumlarda Outlier'lar aşağıdaki gibi analiz edilebilir:

1. Kutu grafiği ile hangi değişkenlerde aykırı değerlerin bulunduğuna dair hızlı bir gözlem yapılması.
2. İlk adımda aykırı değerlere sahip olduğu anlaşılan verilere histogram ile daha yakın bir bakış sağlayarak değişkenin dağılımına dair fikir elde edilmesi.
3. Değişkenin dağılımına göre tek değişkenli istatistiksel bir analiz yöntemi seçilmesi.

Normal dağılımı gözlemlemek için çeyrekler açıklığı metodunu kullanarak aykırı değerlerin sayısına yönelik istatistiksel bir değer elde edelim.



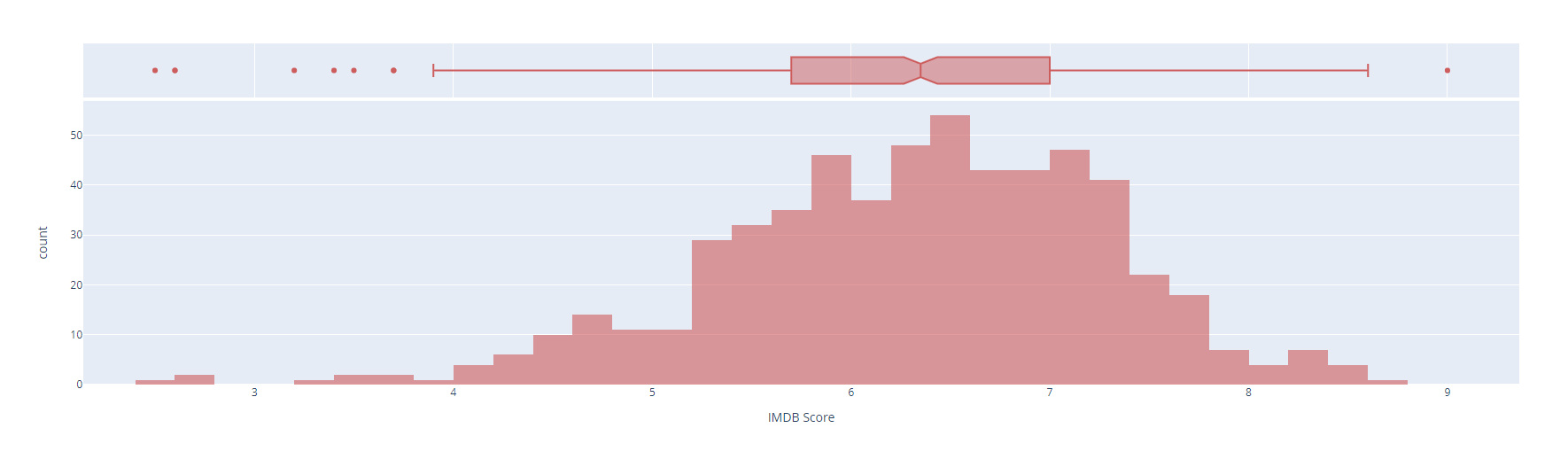
“

“Görsel 1.15 Aykırı değer kutu grafiği”

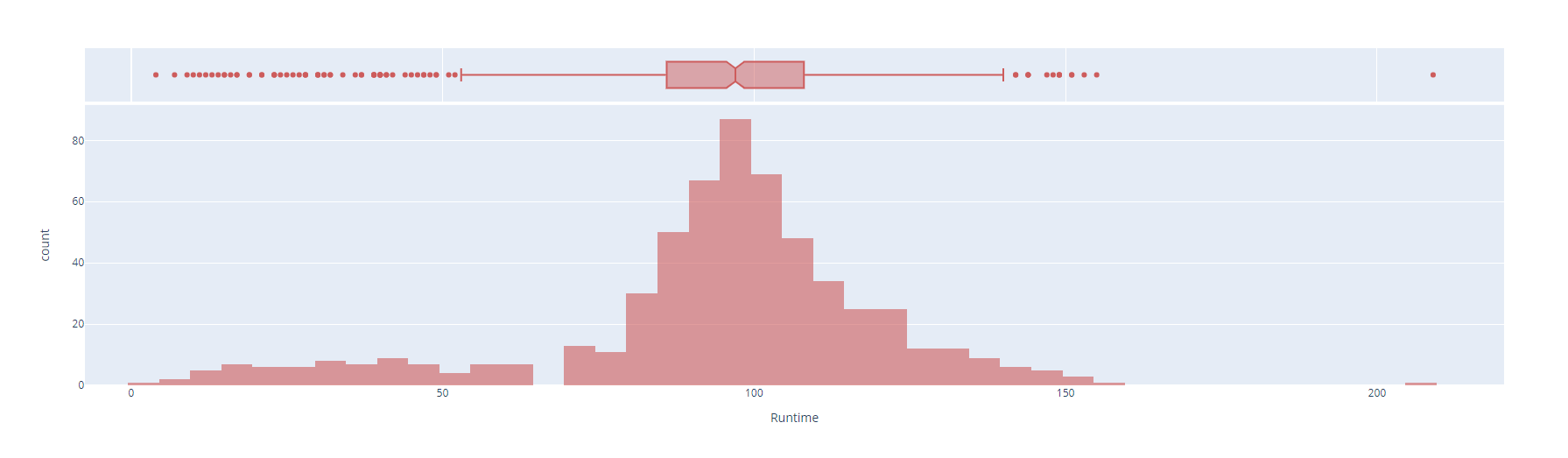
* Histogram Grafiği: Aykırı değerleri istatistiksel olarak da tespit etmeden önce değişkenlerin dağılımına dair fikir edinmemiz gerekiyor. Uygulayacağımız istatistiksel yöntemler normal dağılıma sahip olan ve olmayan değişkenlere göre farklılık göstermek zorundadır.
* Aykırı değere sahip değişkenlerin dağılımına dair fikir elde etmek için histogramdan faydalanacağız.

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

 “Tablo 1.7 IMDB Score-Runtime betimsel değerler”

“IMDB Score Histogram grafiği”



“Runtime-Score Histogram grafiği”

* “IMDB” aykırı değer sayısı: 9
* “Runtime” aykırı değer sayısı: 75
* Sonuç olarak aykırı değerlere karşı daha dayanıklı olan Span of Quarter (IQR) yöntemine göre Runtime değerinde normal dağılımın dışında daha ayrık değerler gözlemleyebiliyoruz.
* “Runtime” değişkeni normal dağılıma aykırı daha fazla veri bulunduruyor.