2019 Dünya Mutluluk Raporu Veri Seti (EDA) Raporu

1. Giriş

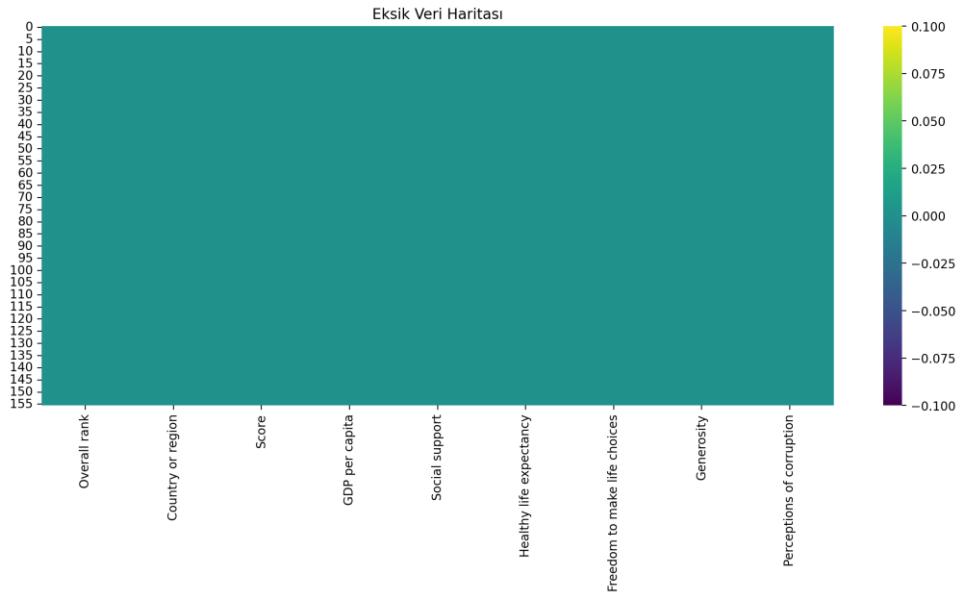
Bu rapor, 2019 Dünya Mutluluk Raporu'ndan alınan veri setinin keşifçi veri analizi (EDA) sonuçlarını sunmaktadır. Amacımız, veri setinin yapısını anlamak, eksik değerleri kontrol etmek, değişken tiplerini incelemek, temel istatistiksel özetler çıkarmak ve görselleştirmeler aracılığıyla her bir özelliğin dağılımını ve aralarındaki ilişkileri ortaya koymaktır.

2. Veri Yükleme ve Genel Bakış

- Veri seti, "2019.csv" dosyasından yüklenmiştir.
- df.head(5): Veri setinin ilk 5 satırına erişildi. Bu, verinin ilk bakışta nasıl göründüğüne dair hızlı bir fikir vermektedir.
- df.info():Veri setinde toplam sütun ve satır bulunmaktadır.
    - Her sütunun veri tipi (object, float64, int64 vb.) ve boş olmayan (non-null) değer sayısı hakkında bilgi alındı. Bu, veri eksikliği veya yanlış veri tipleri gibi olası sorunları erken aşamada tespit etmemizi sağlar.
- df.describe():
    - Sayısal değişkenlerin sayısı, ortalaması, standart sapması, minimum ve maksimum değerleri ile çeyreklikler (25., 50. ve 75. persentil) hakkında temel istatistiksel özet bilgi edinildi. Bu özet, her bir sayısal özelliğin merkezi eğilimini, yayılımını ve olası aykırı değerlerini anlamak için ilk adımdır.

3. Eksik Veri Analizi

- null_control(): Her bir sütundaki boş (null) değer sayıları kontrol edildi. Bu kontrol, veri setinde eksik gözlemlerin olup olmadığını ve hangi sütunlarda yoğunlaştığını gösterir.
- sns.heatmap(df.isnull()) (Eksik Veri Haritası):
    - Seaborn kütüphanesindeki heatmap kullanılarak null (boş) verilerin görsel bir haritası oluşturuldu. Bu harita, veri setindeki eksik verilerin dağılımını ve yoğunluğunu grafiksel olarak görmemizi sağladı. Görsel inceleme sonucunda, veri setinde belirgin bir eksik veri bulunmadığı teyit edildi (eğer görselde boş alanlar olsaydı, oralarda eksik veri olduğu anlaşılırdı).

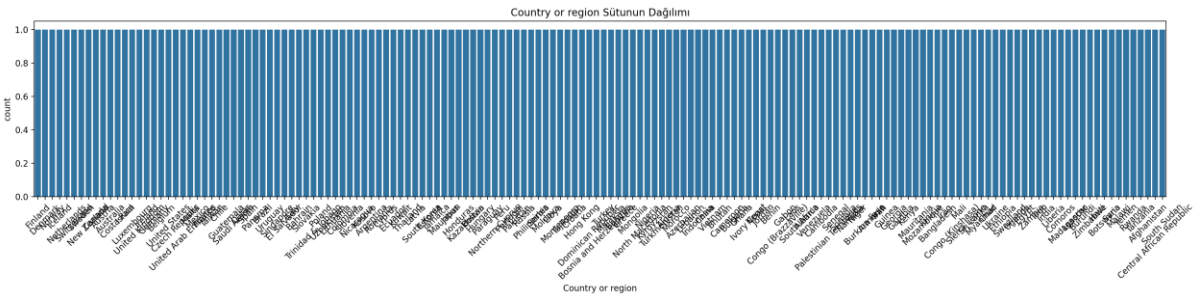


4. Değişken Tipleri ve İstatistiksel Ölçümler

- df.dtypes: DataFrame'deki her sütunun veri tipi listelendi. Bu, veri analizi için doğru yöntemlerin uygulanmasını sağlamak açısından önemlidir.
- Standart Sapma ve Varyans (std_var()): "Score", "GDP per capita", "Social support", "Healthy life expectancy", "Freedom to make life choices", "Generosity", "Perceptions of corruption" sütunlarının standart sapmaları ve varyansları hesaplandı. Bu ölçümler, her bir özelliğin ortalamadan ne kadar saptığını ve verilerin ne kadar dağınık olduğunu gösterir.
- Örneğin, daha yüksek standart sapma ve varyans değerine sahip özellikler, daha geniş bir dağılım sergilemektedir.

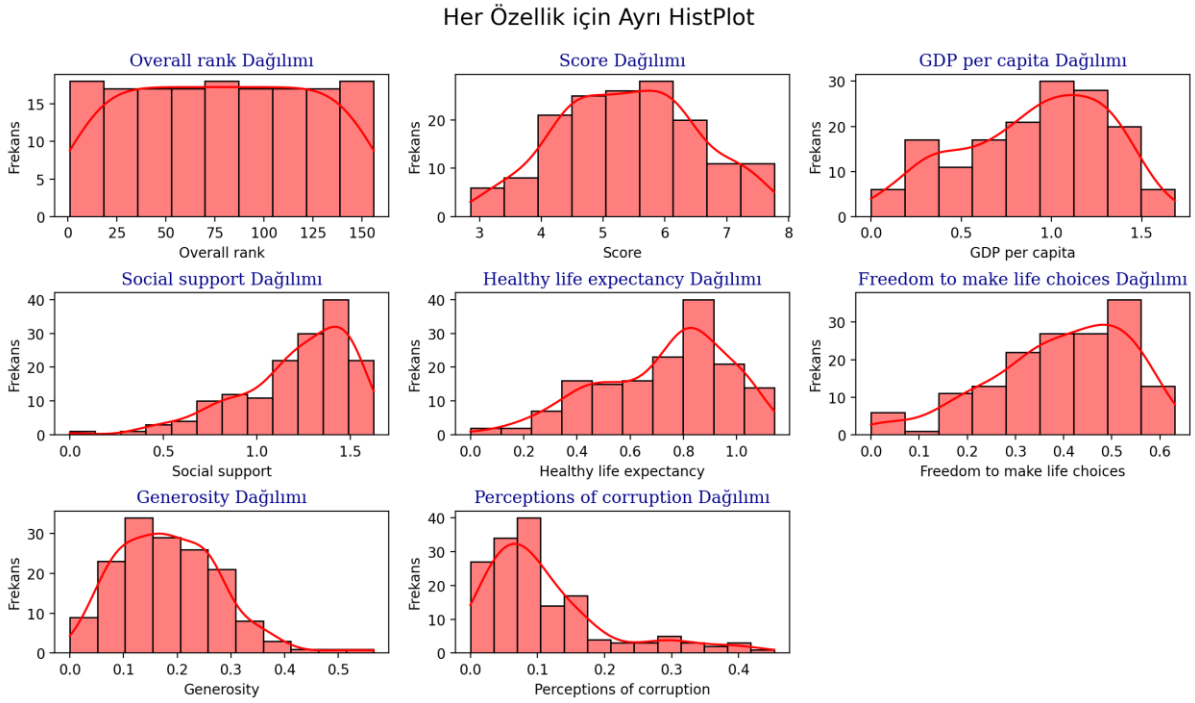
5. Kategorik Değişken Dağılımı

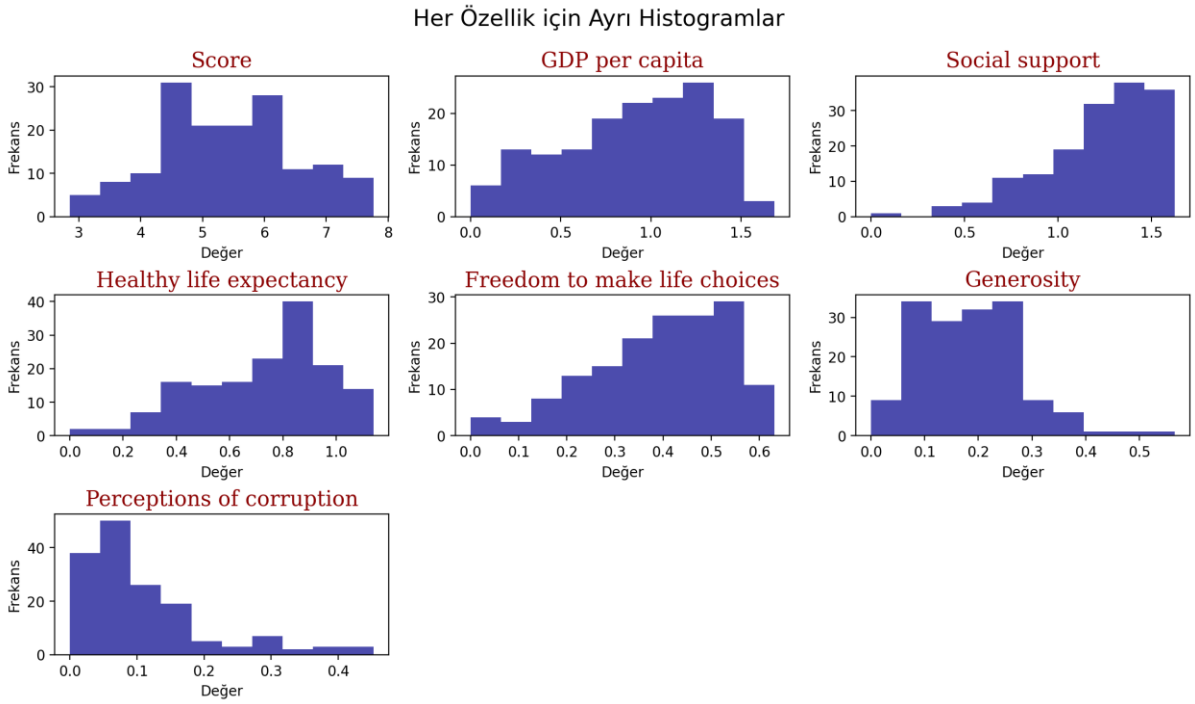
- string_dagilim(): Veri setindeki "object" (kategorik) tipindeki sütunların dağılımı kontrol edildi. value_counts() ile her bir kategorinin frekansları listelendi ve sns.countplot ile görselleştirildi.
- Bu analiz sonucunda, veri setinde yalnızca tek bir kategorik sütun (muhtemelen 'Country or region') olduğu ve her bir kategori değerinin (ülke) yalnızca bir kez geçtiği gözlemlendi. Bu, bu sütunun benzersiz tanımlayıcılar içerdiğini ve doğrudan bir kategori değişkeni olarak kullanılmasından ziyade bir kimlik bilgisi olarak işlev gördüğünü düşündürmektedir.



6. Sayısal Değişken Dağılımları (Histogramlar)

- Veri setindeki sayısal özelliklerin dağılımlarını görselleştirmek için hem hist() (KDE'li) hem de plot_() (temel histogram) fonksiyonları kullanıldı.
- hist() (KDE'li Histogramlar): Bu grafikler, her bir sayısal özelliğin dağılım şeklini (simetri, çarpıklık, mod sayısı) daha net gösteren KDE (Kernel Yoğunluk Tahmini) çizgilerini içerir.
- Score: Yaklaşık simetrik, çan eğrisine benzer bir dağılım.
- GDP per capita, Social support, Healthy life expectancy, Freedom to make life choices: Bu özellikler sola çarpık (negatif çarpık) bir dağılım sergilemektedir. Bu, ilgili değerlerin yüksek kısımlarda yoğunlaştığını ve düşük değerlere doğru daha az sayıda veri noktası olduğunu gösterir. Örneğin, GSYİH'si yüksek olan ülkeler daha yaygındır.
- Generosity, Perceptions of corruption: Bu özellikler belirgin bir şekilde sağa çarpık (pozitif çarpık) bir dağılım sergilemektedir. Bu, ilgili değerlerin düşük kısımlarda yoğunlaştığını ve yüksek değerlere doğru uzanan uzun bir kuyruk olduğunu gösterir. Özellikle "Perceptions of corruption" için yolsuzluk algısının düşük olduğu ülkelerin çok daha yaygın olduğu, yüksek yolsuzluk algısına sahip ülkelerin ise azınlıkta olduğu dikkat çekicidir.
- plot_() (Temel Histogramlar): KDE çizgileri olmadan, yine de dağılım eğilimleri gözlemlenmiştir. Yorumlar, KDE'li histogramlardaki gözlemlerle tutarlıdır, ancak dağılımın incelikleri (örneğin çift modluluk gibi) KDE çizgileri kadar net olmayabilir.
- Özetle, histogramlar veri setindeki özelliklerin çoğunun normal dağılımdan saptığını ve çarpıklık gösterdiğini açıkça ortaya koymaktadır.

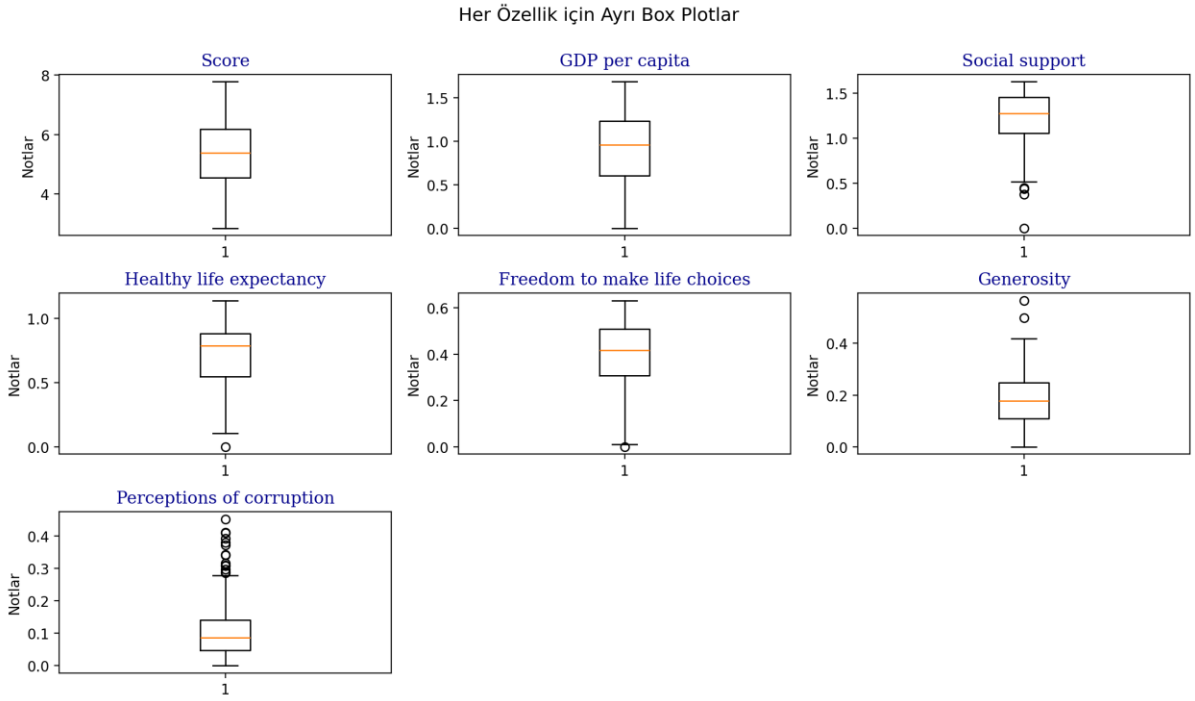




7. Aykırı Değer Analizi (Box Plotlar)

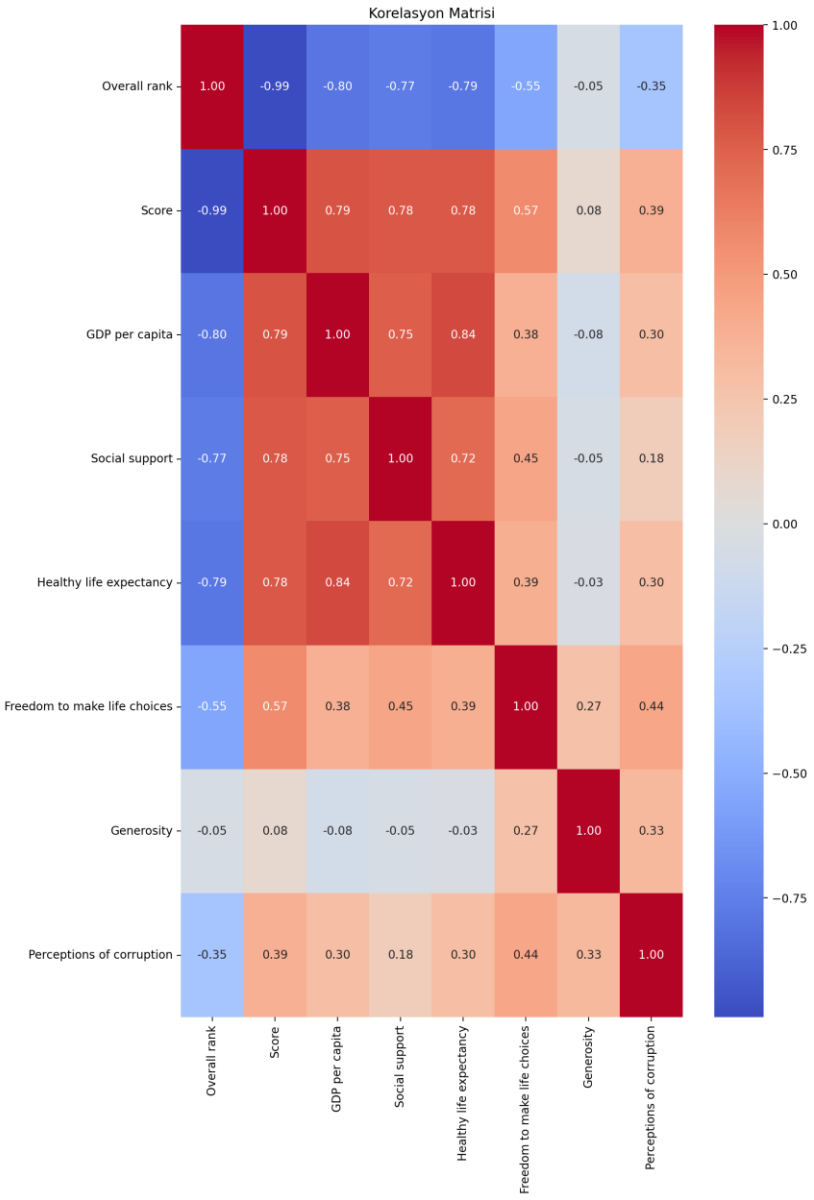
- box() fonksiyonu kullanılarak her bir sayısal özellik için kutu grafikleri (box plot) çizildi. Kutu grafikleri, verinin merkezini, yayılımını ve özellikle aykırı değerleri hızlıca görmemizi sağlar.
- Score, GDP per capita, Freedom to make life choices: Bu özelliklerin kutu grafiklerinde belirgin bir aykırı değer gözlemlenmemiştir. Dağılımları nispeten simetrik veya hafif çarpık görünmektedir.

- Social support ve Healthy life expectancy: Her iki özellikte de düşük değerlerde (kutu ve bıyıkların altında) birkaç aykırı değer bulunmaktadır. Bu, sosyal destek ve sağlıklı yaşam beklentisinin diğer ülkelere göre belirgin şekilde düşük olduğu bazı ülkeler olduğunu gösterir. Bu gözlem, ilgili histogramlardaki sola çarpıklıkla uyumludur.
- Generosity ve Perceptions of corruption: Bu özelliklerde yüksek değerlerde (kutu ve bıyıkların üzerinde) aykırı değerler bulunmaktadır. Özellikle "Perceptions of corruption" için aykırı değerlerin yoğunluğu dikkat çekicidir. Bu, önceki histogramlardaki şiddetli sağa çarpıklıkla tutarlıdır; çoğu ülke düşük yolsuzluk algısına sahipken, az sayıda ülke çok yüksek algılara sahiptir ve bu yüksek algılar aykırı değer olarak görünmektedir.



8. Korelasyon Analizi (Heatmap)

- heat() fonksiyonu kullanılarak sayısal özellikler arasındaki korelasyon matrisi bir ısı haritası (sns.heatmap) ile görselleştirildi.
- Isı haritası, değişkenler arasındaki doğrusal ilişkinin yönünü (pozitif/negatif) ve gücünü gösterir. Koyu renkler güçlü pozitif veya negatif korelasyonları, açık renkler ise zayıf korelasyonları işaret eder.
- Bu görselleştirme, hangi özelliklerin birbirleriyle güçlü bir şekilde ilişkili olduğunu hızlıca görmemizi sağlar. Örneğin, "Score"un diğer bazı faktörlerle (örn. GDP per capita, Social support) güçlü pozitif korelasyonlar göstermesi beklenir.



9. Sonuç ve Gelecek Adımlar

- Bu keşifçi veri analizi raporu, 2019 Dünya Mutluluk Raporu veri setinin temel özelliklerini kapsamlı bir şekilde incelemiştir. Veri tipleri, eksik değerler, temel istatistikler ve her bir özelliğin dağılım şekli hakkında önemli bilgiler elde edilmiştir.
  Ana Çıkarımlar:
- Veri setinde belirgin bir eksik değer bulunmamaktadır.
- Birçok sayısal özellik normal dağılımdan sapma göstermekte (sağa veya sola çarpık).
- Özellikle "Social support", "Healthy life expectancy", "Generosity" ve "Perceptions of corruption" özelliklerinde dikkat çekici aykırı değerler tespit edilmiştir. Bu aykırı değerler, sonraki analizlerde (örneğin modelleme) özel işlem gerektirebilir (örneğin aykırı değer tedavisi veya robust yöntemler).

- Korelasyon analizi, değişkenler arasındaki ilişkileri anlamak için bir başlangıç noktası sağlamıştır.
- Bu bulgular, veri ön işleme, özellik mühendisliği ve makine öğrenimi modellerinin geliştirilmesi gibi sonraki adımlar için değerli bir temel oluşturmaktadır. Örneğin, çarpık dağılımlar için veri dönüşümleri düşünülebilir veya aykırı değerlerin etkisini azaltmak için stratejiler uygulanabilir.