COVID-19 VERİ ANALİZİ

Furkan PAŞAOĞLU 160201056

Bilgisayar Mühendisliği Bölümü

Kocaeli Üniversitesi

[furkan\_pasaoglu@hotmail.com](mailto:furkan_pasaoglu@hotmail.com)

Özet

*WHO(World Health Orgination) tarafından küresel salgın olarak ilan edilen Yeni Korona Virüs hastalığı (COVID-19) üzerinde veri analizi ve görselleştirme yöntemleri kullanırak büyük verinin analizi farklı platformlarda yapılmıştır. Python ve veri analizi için oluşturulmuş kütüphaneleri Pandas,Numpy - veri grafiklendirilmesi için oluşturulmuş Matplotlib, Seaborn ve Plotly teknolojilerinden yararlanıldı.Bu analizler Jupyter Notebook üzerinde yazılmıştır.*

*Ayrıca veriler üzerinde SQLite kullanılarak aranan sorulara SQL üzerinden cevap bulunmuştur.SQLite kernel kurulumu yapılarak SQLite’ın Jupyter Notebook üzerinde çalışması sağlanmıştır [1].*

*COVID-19 verileri Tableau platformu kullanılarak veri görselleştirme işlemleri bu platformda da yapılmış ve platform içerisinde bulunan DashBoard ile analiz edilen ve görselleştirilen veriler birleştirilmiştir.*

*Keyword : Covid-19, Veri Analizi , Pandas , SQL , Tableau*

# Giriş

Dünya tarihinde eşi benzeri görülmemiş bir olaya tanıklık ediyoruz. Yeni Nesil Korona Virüs(Novel Corona Virus) adı konulan ve Dünya’da hızla yayılan korona virüs durdurak bilmiyor.Bu çalışmamızda kaynak [2] ‘ de belirtilen veri seti kullanılmıştır.Bu veri seti korona virüsün ilk çıktığı günden bugüne her gün vaka , ölüm ve kurtatılan kişi sayısını analiz etmemizi sağlayan ve her gün güncellenen bir veri setidir.

Bu çalışmamızda Korona Virüs analizi ve görselleştirilmesi ile ilgili 10 adet soru, 9 Python, 1’i SQL ve 1 adet Tableau uygulamamız mevcuttur. Bu projeyi yapmamızda ki amaç Yeni Nesil Korona Virüsün ne denli hızla yayıldığını somut bir şekilde göstermek , insanoğlunun bilgisayarsız günlerini alacağı çalışmaları bilgisayar ve yazılım yardımı ile saniyeler içerisinde analiz etmek . Bu analizlerin daha sonradan Yapay Zeka teknolojileri insanoğlunda bu zorlu süreçte tedevi aşamasında yardımcı olmaktır.

Korona Virüs veri analizi çalışmaları literatürde mevcuttur.Kaggle’ın yayınlamış olduğu veri seti ve ödüllü yarışması bu analize olan talebi arttırmıştır.

# Temel Bilgiler

Projeyi gerçekleştirirken Anaconda Platformu , bu platform içerisinde veri analizine dair kütüphaneleri (Python , Pandas, Numy , Matplotlib , Seaborn) , çalışma ortamımız Jupyter Notebook’ içerisinde bulundurduğu için bu teknoloji kullanılmıştır.

Plotly kütüphanesi Anaconda üzerine harici olarak kuruldu ve kullanıldı.

Tableau platformu kullanıldı, veri görselleştirme için en popüler platformlardandır.

SQL motoru olan , SQLite teknolojisi kullanıldı.

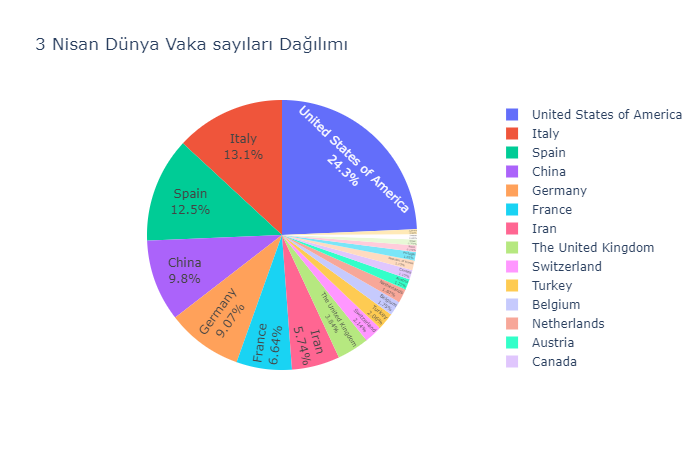
# Yöntem

COVID-19 veri analizi için 10 adet soru ve 1 adet Tableau uygulamamız mevcuttur.Bu soruların ve uygulamanın ne olduğunu ve nasıl yapıldığını inceleyim.

İlk olarak kütüphaneleri import ederek başlıyoruz.Ardından veri setimizi github linkinden çekiyoruz[3].

# Soru 1 : 3 Nisan tarihine ait vaka sayısı dağılımı ve görselleştirilmesi

Bu cevap için ilk olarak veri setimizde bulunan “who\_covid\_19\_sit\_rep\_time\_series.csv” dosyasının github’tan çekiyor ve işleme sokuyoruz. Pandas.Pivot\_table kullanarak Ülkeye göre gruplandırıyor ve 4/3/20 tarihinde işlem yapacağımız için aggeregation olarak ‘sum’ fonksiyonun aktifleştiriyor . Plotly ile pie şeklinde çizim yaptırıyoruz.

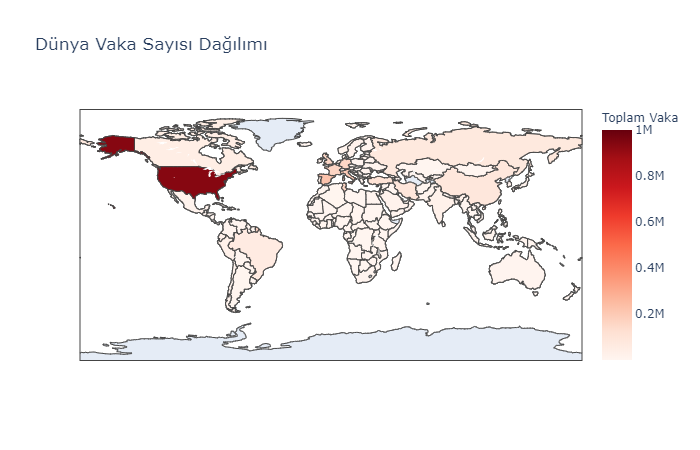


**Soru2 :**Dünyada görülen vaka sayılarının Dünya haritası üzerinde analizi

**Cevap2 :** Vaka sayılarının bulunduğu bir diğer csv dosyası olan time\_series\_covid19\_confirmed\_global kullanılarak confirm adında vaka sayılarının bulunduğu dosya githubtan okundu.

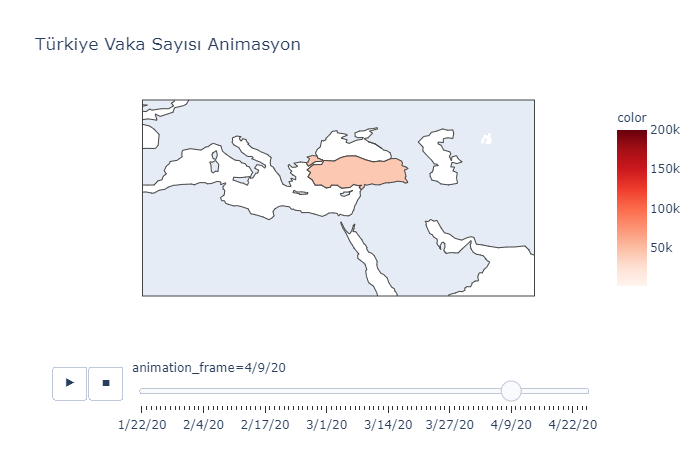
Ardından her ülkeye ait bir enlem ve boylam bilgisini kullanmak için ilk olarak Long ve Son kolon bilgisini pivot\_table ile birleştirildi.Daha sonradan o Long değerine ait Lat değeri çekilerek ve **drop\_duplicate** komutu kullanılarak her ülkeye ait enlem , boylam bilgileri bulundu.

Daha sonra plotly.choropleth fonksiyonunda gerekli parametleri vererek dünya haritası çıkarmış olduk.



**Soru 3 :** Türkiye'nin vaka sayılarının gün gün harita üzerinde animasyon olarak gösterilmesi

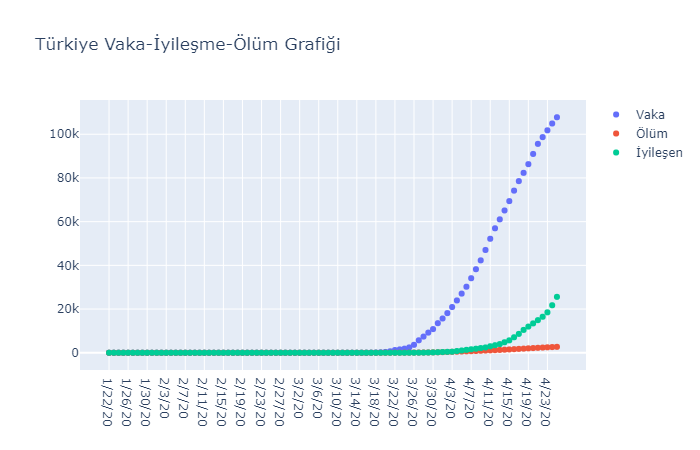
**Cevap 3 :** plotly içerisinde animasyon olarak gösterim sağlayan fonksiyon animasyon olarak gösterilecek veri boyutu kadar ülke adı istemekte , örneğin 12 gün için 12 ülke adı istemekte . Bundan dolayı kodumuzda ilk olarak veri boyutumuzun uzunluğunu len() fonksiyonu ile hesapladık ve içerisinde Ülke , Lat, Long , State bilgileri içerdiği için sayımızdan bu 4 maddeyide düşürdük . İlk olarak kayıtlı bir ülke geleceğinden oradan da 1 sayı daha düşürdük. Toplam gün sayımız day\_count-5 olacak şekilde ayarlandı. Bir for döngüsü ile kopyalama işleminden geçirdik. Böylece gün sayısı kadar kopya oluşturulmuş olacaktı.

Ardından plotly.choropleth fonksiyonu içerisine tanımladığımzı animation\_frame ile animasyon gösterimi tamamlandı.Günbegün artan vaka sayısı kırmızıdan siyaha artan renkler ile gösterildi

**Soru 4 :** Türkiye Durum Analizi ve Grafiklendirilmesi

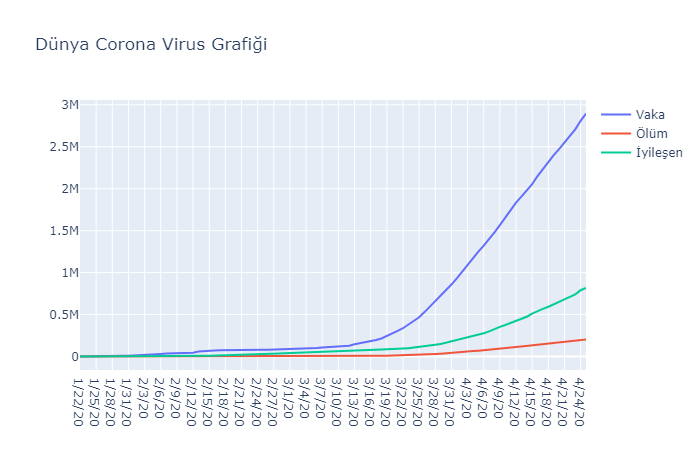
**Cevap 4 :** Confirm dataframe üzerinde sadece Türkiye vaka bilgisi bulunduğu Turkey, Death dataframe üzerinde sadece Türkiye ölü sayısının ve bilgisi bulunduğu TurkeyDeath, Recovered dataframe üzerinde sadece Türkiye ülkesinin kurtarılma sayısının bulunduğu TurkeyRecov matrisleri oluşturuldu.Daha sonradan bu matrisler üzerinde işlem yapabilmek adına Transpozu alındı.

Görselleştirilmesi için index bilgileri (gün bilgileri) x eksenine, values bilgileri (sayı) y eksenine eklendi ve values bilgileri , index bilgilerine göre reshape işleminde geçti.

Ardından plotly.go kullanılarak birleştirilen durum bilgileri ekrana yansıtıldı.

**Soru 5 :**Dünya Durum Analizi ve Grafiklendirilmesi

**Cevap 5 :** Confirm , Death, Recovered dataframeler üzerinde iloc işlemi ile sadece gün kolonların bulunduğu 4.kolondan son kolona kadar sum() fonksiyonu ile toplama işlemi yapıldı. Böylece günbegün vaka , ölü , kurtatıralan sayılarının bulunduğu dataframe’ler oluşturuldu. Ardından Soru 4 ‘te yapıldığı gibi index ve value değerleri plotly.go fonksiyonuna parametre olarak verildi ve grafik çizdirildi



**Soru 6 :** Ülkelerin güncel Vaka-Ölüm-Kurtarılan sayılarının belirlenmesi ve Vaka Sayısına göre sıralanması

**Cevap 6 :** Vaka – Ölüm – Kurtarılan sayılarını bulmak için ilk olarak Confirm tablosundan 1. kolon ve -1. yani son kolon bilgileri alınır, bu kolonlar Ülke ve Son Vaka Sayısı toplamıdır. Daha sonra groupby ile ülkelere göre gruplandırıp sum() ile topluyoruz. Böylece vaka sayılarını buluyoruz. Aynı şekilde Death ve Recovered dataframe’leri için bu durumu yapıp ardından bu tabloluları birleştirme işlemine geçiyoruz.

Pandas.merge ile iki tablo birleştirilebileceğinden ilk 2 tabloyu alıyor Ülke adına göre birleştiriyor ve **genel** adlı değişkene atama yapıyoruz . Daha sonradan son tablo ile genel tablosunu birleştirip **genel** tablosuna atama yapıoruz.

Böylece 3 bilginin aynı ayna yer aldığı genel tablosu elde ettik.

Genel tablosunu **sort\_values** fonksiyonu ile Vaka sayısına göre gruplandırdık

**Soru 7** : Dünyada vaka-ölüm-kurtarılma oranlarının hesaplanması

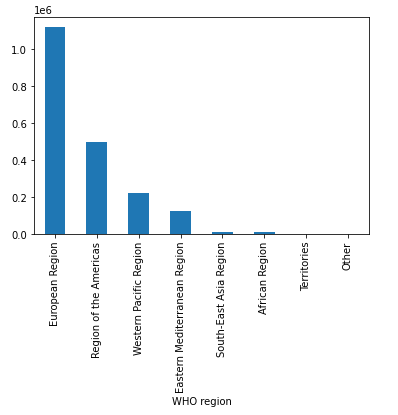
**Cevap 7 :** Soru6 çözerken elde etmiş olduğumuz **genel** dataframe aynı şekilde bu sorunun çözümünde de kullanılacaktır.Genel tablosunda sum() fonksiyonunu çalıştırarak Dünya üzerinde toplam sayıları görebiliyor olduk.Ardından bu sayıları toplam Dünya nüfusü olan 7.8 Milyara bölüp 10 bin ile çarpıyoruz.

Bu bilgiler bize her 10.000 kişide virüsün durumunu gösteriyor.



**Soru 8** : Hastalıktan en çok etkilenen kıtaların sıralanması

**Cevap 8 :** İlk olarak time\_series dataframe üzerinden”WHO Region” kolonun göre gruplandırma işlemini yapıyoruz. Böylece kıtaları gruplandırmış olduk, sum() fonksiyonu ile her bir kıtada vaka görülme sayısını elde ettik. Matplotlib kütüphanesinden plot ile çizim sağladık.



**Soru 9.1 :** Amerika , İtalya ve Çin Devletinin Vaka Sayısının karşılaştırılması

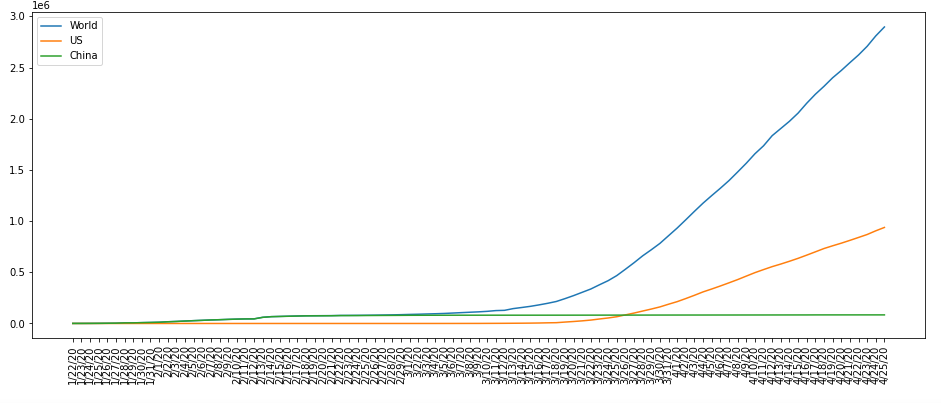
**Soru 9.2 :** Dünya, Amerika ve Çin Vaka Sayısının karşılaştırılması

**Cevap 9 :** Sorumuz iki şıktan oluşmakta ama cevapları benzer olduğu için tek cevaba sığdırık. Vaka sayılarını karşılaştırmak için ilk olarak confirm dataframe’i ülkelere göre gruplandırıp , sum() fonksiyonu ile topladık.

Ardından istenilen ülke bilgisine göre **loc** işlemi yapıldı.

Örneğin; US = countries.loc["US",:] işlemi ile Amerika Vaka sayısı bulundu.Aynı şekilde karşılaştırılacak ülkelerinde loc işlemi yapılır.

Plot ile çizdirilir.



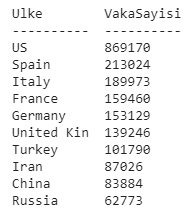
**Soru 10 :** SQL üzerinde yapılan analiz

.mode csv

.import "series\_covid19\_confirmed\_global.csv" confirm

komutları kullanılarak tablo okunur.

**Soru 10.1 :** En fazla vaka görülen ilk 10 ülke

**Cevap 10.1 :** Tablodan ülkeyi ve sum(Gün) ile Vaka sayısını select ediyoruz. confirm tablosundan from ediyor, ülkere göre gruplandırıp , Vaka sayisina göre order by DESC , limit 10 ile ilk 10 satırı getiriyoruz.

**Soru 10.2 :** Bugün en fazla yeni vaka tespit edilen ilk 10 ülke

**Cevap 10.2 :** Yeni Vakayı tespit etmek için bugün ile beraber toplam vaka sayısından dün itibariyle olan toplam vaka sayısı çıkartılır.Yeni vakaya göre sıralama yapılır.



**Soru 10.3 :** Nüfus sayısına oranla en fazla vaka sayısına sahip ilk 10 ülke

Cevap 10.3 : Bu sorunun çözümü için nüfus bilgilerinin olduğu yeni bir verisetine ihtiyaç vardır. Bu veri seti [4]’te belirtilen kaynaktan indirildi.

İmport ile populationCountry olarak adlandırılan bir tablo oluşturuldu.Tabloda US adı United States olarak tanımlı olduğu için Update komutu ile US adına güncellendi.

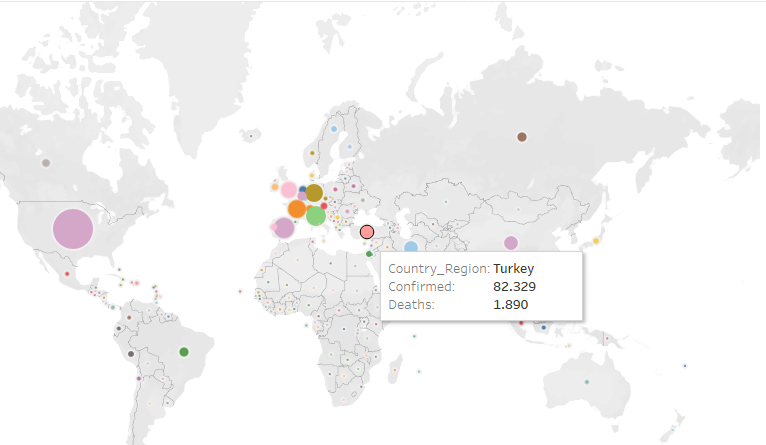
INNER JOIN işlemi kullanılarak tablo birleştirildi[5].Değerler çok küçük çıktığından dolayı nüfus 1 milyona bölündü ve o Ülkede milyon başına düşen ortalama vaka sayısı bulundu.



San Marino ülkesinde her 1milyonda 14.771 kişi korona virüsüne yakalanmış durumda.

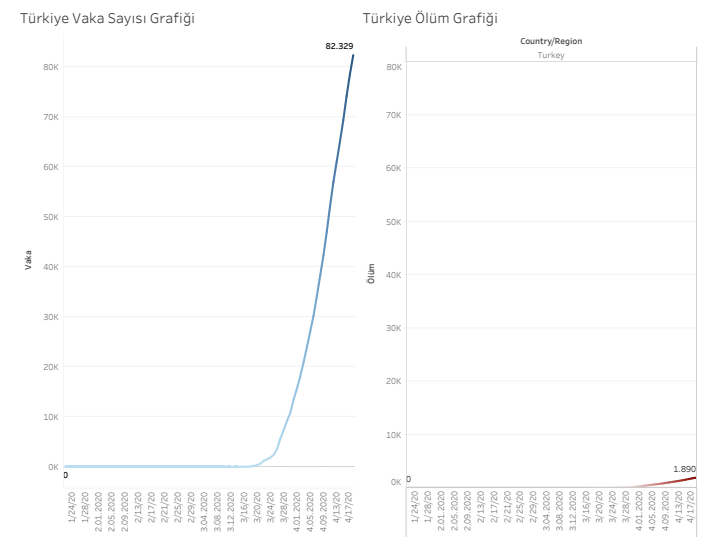
**Soru 11 :Tableau Çalışması :**

**11.1.**



Bu çalışma enlem ve boylamların atanması , renklerin ülkelere göre belirlenmesi ve Boyutun ise Vaka sayısına göre belirlenmesi sonucu oluşmuştur.

**11.2.**

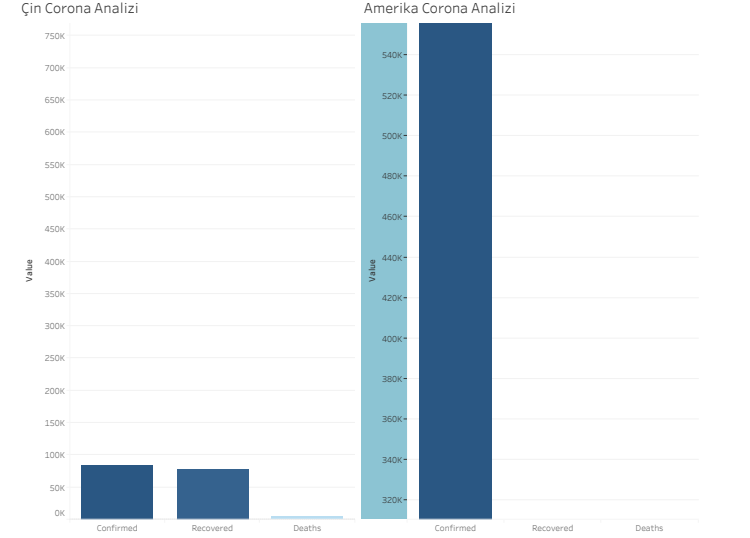


Filtreleme ile Türkiye ülkesinin seçildiği ve x ekseni tarih, y ekseninde sayıların olduğu bir tablo oluşturuldu.

Aynı şekilde oluşturulan ölüm sayısı grafiği

Tableau dashboard üzerinde birleştirildi[6].

# 11.3.



Bu çalışmaya ilk olarak filtreleme ile ülkenin belirlenmesi ile başlandı.Ardından günbegün vakasayılarının bulunmuş olduğu tablonun tümü toplanır.Aynı şekilde ölüm ve kurtarılma sayılarıda toplanıp , Bar chart şeklinde çizim yapılır[7].

Çin ve Amerika için oluşturulan grafikler, dashboard ile birleştirildi.

# Kaynakça

1. <https://github.com/brownan/sqlite3-kernel>
2. <https://github.com/CSSEGISandData/COVID-19>
3. <https://stackoverflow.com/questions/55240330/how-to-read-csv-file-from-github-using-pandas>
4. <https://www.kaggle.com/tanuprabhu/population-by-country-2020>
5. <https://www.sqlitetutorial.net/sqlite-inner-join/>
6. <https://www.tableau.com/learn/articles/business-intelligence-dashboards-examples>
7. <https://help.tableau.com/current/pro/desktop/en-us/buildexamples_bar.htm>