**KOCAELİ ÜNİVERSİTESİ**

**BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ**

**BÜYÜK VERİ ANALİZİNE GİRİŞ DERSİ VİZE ÖDEVİ**

COVID-19 VİRÜS SALGINI ANALİZ VE GÖRSELLEŞTİRMESİ

MURAT ALTINAY 170202109

170202109@kocaeli.edu.tr

# giris

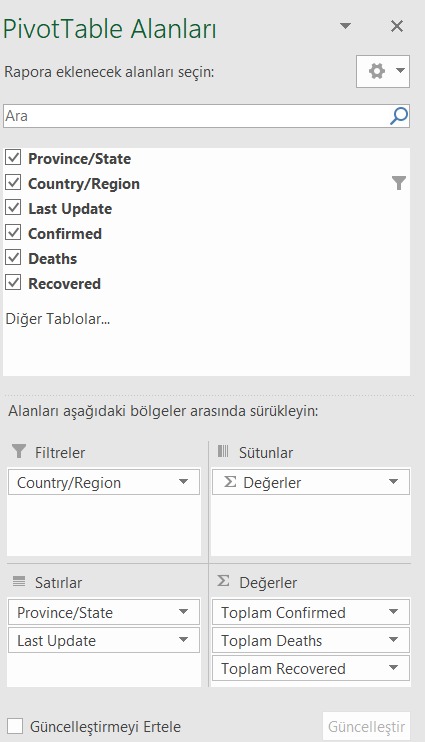
Öncelikle dünya çapıda yayılan bu salgını anlamak ve analiz etmek için günlük kayıtları tutulan veriler birleştirildi ve belirli problemlere çözümler oluşturuldu.

# Temel bilgiler

Görselleştirme ve analiz işlemleri için python dili, matplotlib ve seaborn kütüphaneleri kullanılmıştır.

# Tasarım

* Öncelikle kaynakçada belirtilen veri seti indirilmiştir.[1]
  + İndirilen veri seti incelenmiştir ve covid-19 günlük rapor klasörü içerisindeki csv dosyaları birleştirilmiştir.
  + Birleştirme işlemi için cmd komut satırında ilgili dizine gidip **‘copy \*.csv yenibelge.csv’** komutu yazılır.Bu sayede yenibelge.csv dosyasına klasör içerisindeki büyün csv dosyaları kopyalanmış olur.
* Veriler kümülatif artış göstermektedir.Fakat US verileri gün içinde birden fazla olduğu için işlem yapmayı kolaylaştırmak adına tablo pivot tablosuna eklenerek US için filtrelenerek aynı gün verileri toplanmıştır.



* Daha sonra excellden daha fazla yararlanabilmek adına bu kümülatif artan verilerden günlük vaka artışını bulmak için fonksiyon yazılmıştır.
* =EĞER(A2=A1,D2-D1,D2)
  + Yazılan fonksiyon eyaletleri aynı olan yerlerin bir gün öncesindeki vaka sayısını o günden çıkararak günlük artış değerini yeni bir sütuna yazar.
  + Bir hücre için yazılan bu fonksiyon fareyle tıklayıp kaydırma işlemi yaparak bütün hücreler için uygulanmış oldu.
  + Aynı fonksiyon daha sonra uygun filtreleme yapılarak ülkesi aynı olan yerler için de yazıldı ve diğer işlemler de yapıldı.
  + Bu işlemi ikiye ayırmamızın sebebi bazı ülkeler için şehir şehir veri girişi yapılmışken bazı ülkeler için sadece günlük toplam değerler verilmiştir.
  + Bu yüzden önce şehir şehir verilen ülkeler için filtreleme işlemi yapılıp şehir sütunu adan zye sıralanır.Daha sonra bahsedilen fonksiyon yazılır ve son şehir verisine kadar kaydırma işlemi yapılır.Bu noktadan sonrası için yeni fonksiyon yazılır.Fonksiyon için bu sefer şehir sütunu değil ülke sütunu kullanılır .Sonrasında bir hücre için yazılmış bu fonksiyon fareyle aşağı kadar kaydırılarak bütün satırlar için günlük onaylanmış veri sütunu tamamlanmış olur.

# Tasarım aşamaları

* Öncelikle Jüpiter kullanılarak python 3 dosyası oluşturuldu.Daha sonra csv dosyaları pandas kütüphanesi yardımıyla data frame formatında okundu.Gerekli kütüphaneler import edildi.
* Daha sonra problemler oluşturarak çözüm arandı.

#### **Problem 1**: 23 Nisan tarihi itibariyle ülkelere göre onaylanmış vaka sayılarını büyükten küçüğe sıralayınız.

* Problem çözümü için öncelikle veriler 23 nisan tarihine kadar olan toplam vaka sayılarını bulmamız gerekiyor.Veriler kümülatif olarak arttığı için 23 nisan tarihine göre işimize yarayacak sütunlardaki verileri data frame den yeni bir değişkene atama işlemi gerçekleştirilir.
* df\_sub=df[['Country/Region','Last\_Update','Confirmed']]
* df\_sub=df\_sub[df\_sub['Last\_Update']=='2020-04-23'].groupby(['Last\_Update','Country/Region']).sum()

Daha sonra vaka sayılarına göre sıralama işlemi yapılır.İlk on ülke görselleştirilir.Bu soru için görselleştirme işlemi için seaborn kütüphanesi kullanıldı ve bar grafik tercih edildi.

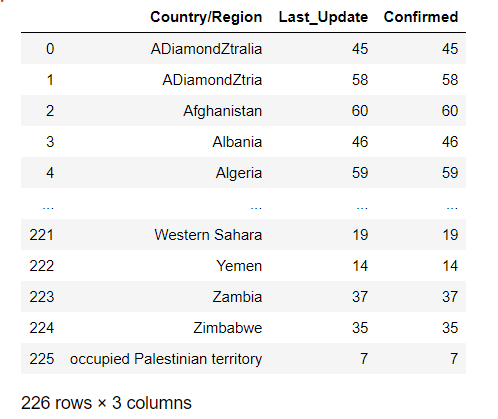
#### **Problem 2:** Covid-19 virüsün ülkelere göre bir günde ortalama yayılma hızlarını karşılaştırınız.

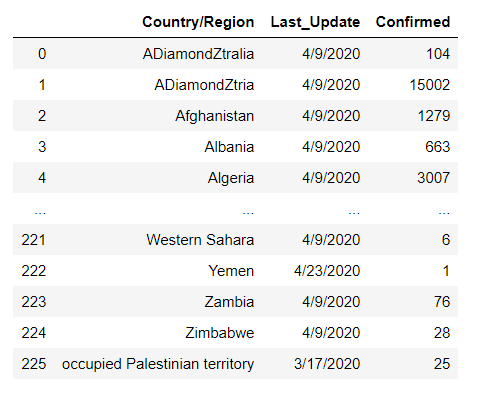
* Virüsün herhangi bir ülkede görüldüğü ilk gün ve 23 Nisan tarihleri arasındaki günlük ortalama yeni vaka sayılarına göre yayılma hızları karşılaştırılmıştır.Bu işlem için aşağıdaki metodlar kullanılmıştır.
* dff=df\_sub.groupby(['Country/Region']).agg('count')

dff.reset\_index(inplace=True)

* dff1=df\_sub.groupby(['Country/Region']).agg('max')
* agg(‘count’) sayesinde bir ülkeye ait kaç günlük veri olduğu saydırılmış olur.
* agg(‘max’) sayesinde ise veriler vaka sayıları kümülatif verildiği için en yüksek sayıya ulaştığı vaka sayısı bulunmuş olur.
* Toplam vaka sayısı da toplam gün sayısına bölündüğünde her ülke için günlük ortalama vaka artış sayısı bulunmuş olur.
* Görselleştirme için en yüksek 15 ülke baz alınmıştır.head(15) metodu kullanılmıştır.

Problem 2 için yazılan kodlardaki dff ve ddf 1 dataframe çıktıları.:





#### **Problem 3:** Günlük ortalama yayılma hızı en yüksek ilk beş ülkenin vaka artış grafiklerinin grafikle karşılaştırılması.

* Önceki soruda bulunan en yüksek yayılma hızına sahip beş ülkenin yayılma hızlarını günlük olarak aynı grafikte gösterilmiştir.Görselleştirme işlemi için matplotlib kütüphanesi kullanılmıştır.

#### **Problem 4:** Türkiye'deki ölüm oranını zamana göre grafiğini görselleştiriniz.

* Dataframe Türkiye verilerine göre filtrelenir günlük ölüm oranı hesaplanır ve görselleştirilir.Ölüm oranı hayatını kaybeden insan sayısının hastalığa yakalanan insan sayısına bölünmesiyle elde edilir.

#### **Problem 5:** En çok nüfus popülasyonuna sahip ilk beş ülkenin günkük vaka artış sayılarının grafik ile karşılaştırılması.

* Bu işlem için dataset içerisinde ülkelerin nüfüus bilgisi olan csv dosyası kullanılmıştır. Bu csv dosysı uygun filtreleme işlemi yapılarak sadece ülke isimleri ve nüfusları olacak şekilde farklı bir csv dosyasına aktarılmıştır ve bu dosya pandas yardımıyla okunmuştur.
* Populasyon verileri sıralanır ve ilk beş ülke verileri dataframe içerisinde aranır ve grafik halinde karşılaştırılır.

#### **Problem 6:** 23 Nisan tarihi itibariyle ülkelerin iyileşen insan sayılarına göre grafikle karşılaştırılması.

* En çok iyileşme sayısına ulaşmış ilk on beş ülke baz alınarak oluşturulmuştur.

#### **Problem 7:**Amerika Birleşik Devletlerinde Covid-19 virisünün yüzde olarak artışını grafik ile gösteriniz.

* Günlük yeni vaka sayılarının mevcut toplam vaka sayısına bölünmesiyle elde edilen zamana göre artış grafiği oluşturuldu.

#### **Problem 8:** Amerika Birleşik Devletlerinde günlük test yapılan yapılan kişilerin yüzde kaçında hastalık tespit edilmiştir?

* Amerikaya özel csv dosyası kullanılarak oluşturulmuştur 12-24 Nisan arasındaki 12 günlük veriler baz alınmıştır.

#### **Problem 9:** Amerika Birleşik Devletleri içerisinde 23 Nisan itibariyle şehirlere göre onaylanmış toplam vaka sayısını bar grafikle gösteriniz.

* states=df[df['Country/Region']=='US']
* states=states[states['Last\_Update']=='2020-04-23']
* states=states[['Province/State','Confirmed']]
* states.plot(x='Province/State',y='Confirmed',kind='bar',color='orange')
* Data Frame verileri US verilerine göre çekilir ve 23 nisan günü yani toplam vaka sayılarını alabilecek şekilde filtrelenir.Daha sonra plot fonksiyonu sayesinde şehir isimleri ve vaka sayıları gösterilir.

#### **Problem 10:** 23 Nisan itibariyle dünya çapında hasta ,kurtulan ve hayatının kaybeden toplam insan sayılarını gösteriniz.

* df\_sub=df[df['Last\_Update']=='2020-04-23'].groupby(['Last\_Update','Country/Region']).sum()
* df\_sub.reset\_index(inplace=True)
* plt.figure(figsize=(10,6))
* df\_sub[['Confirmed','Deaths','Recovered']].agg('sum').plot(kind='bar')
* Tarih ve ülke şeklinde gruplama işlemi yapılır.Bu sayede şehir bilgisine sahip veriler olsa bile sum() komutu ile vaka sayıları ülkeye göre toplanmış olur.
* Yani nihayetinde elimizde gün gün ülke ülke toplam vaka sayıları elde edilmiş olur.

# kazanımlar

Python becerileri geliştirmemi sağladı ve artık verileri excel ve python kullanarak analiz edebiliyorum.Analiz ettiğim verileri anlamlandırmak için görselleştirme işlemleri yapabiliyorum.

##### Kaynaklar

**[1] https://github.com/CSSEGISandData/COVID-19**

[2]<https://seaborn.pydata.org/generated/seaborn.barplot.html>

[3] <https://matplotlib.org/>