FDA Final Project Report

第11組 劉松霖 陳品修 張君豪

1. **Dataset:**

選擇**COVID-19 Dataset**，裡面包含:

1. covid\_19\_clean\_complete.csv
2. tests.csv
3. usa\_county\_wise.csv。
4. **目標:**

分析Data Set，並以regression預測死亡率，以LSTM時間序列預測現有病人數。

1. **分析:**
2. covid\_19\_clean\_complete.csv

看前三列資料:

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

欄位有Province/State、Country/Region、Lat、Long、Date、Confirmed、 Deaths、Recovered

一張含有 螢幕擷取畫面 的圖片

自動產生的描述←看各欄位，只有Province/State有大量空格。

計算目前病人數  
一張含有 螢幕擷取畫面 的圖片

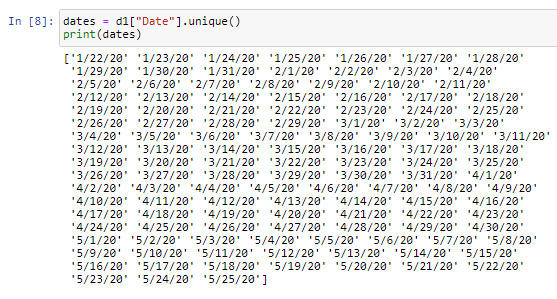
自動產生的描述

列出所有國家:

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

列出日期範圍:



＊有部分國家的資料紀錄方式不同，同天會有不同地區的累積資料，在依國家分類作圖時需要進行修正

需要修正的國家有:

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

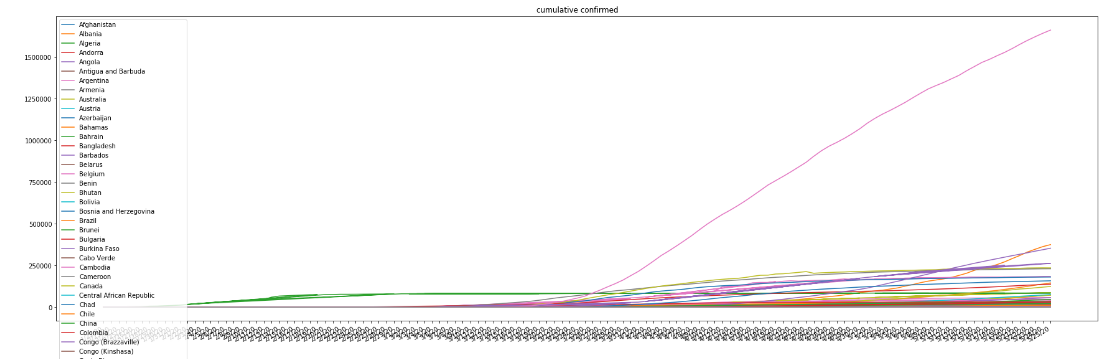
進行修正:

一張含有 螢幕擷取畫面, 鳥 的圖片

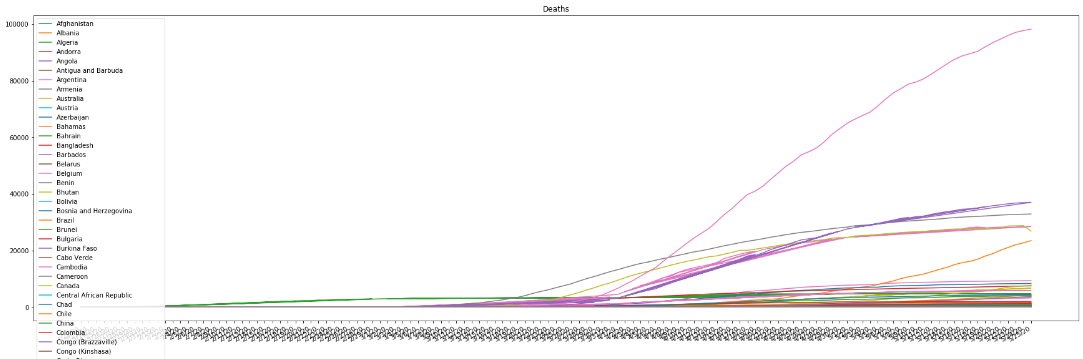
自動產生的描述



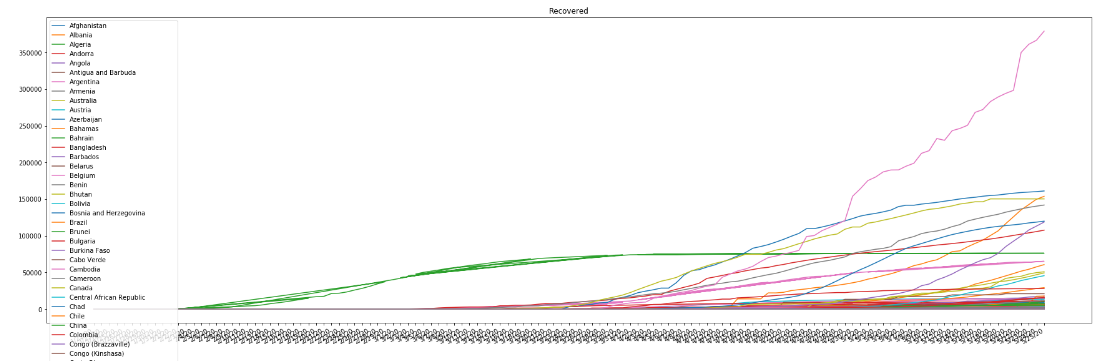
對確診人數作圖:



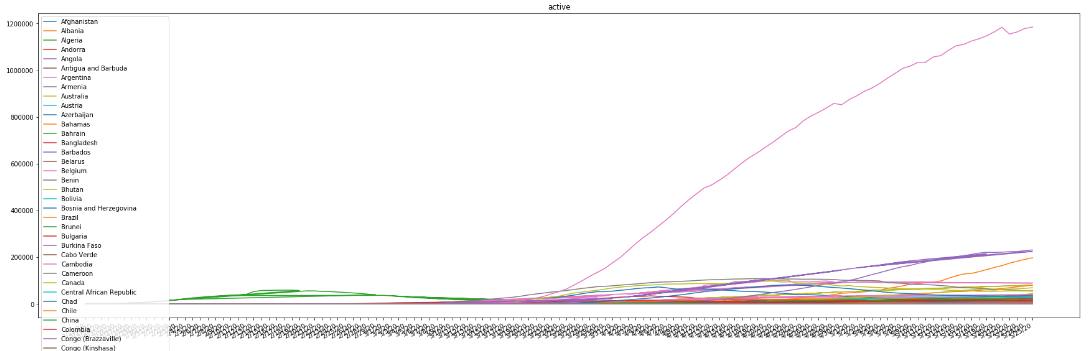
對死亡人數作圖:



對康復人數作圖:



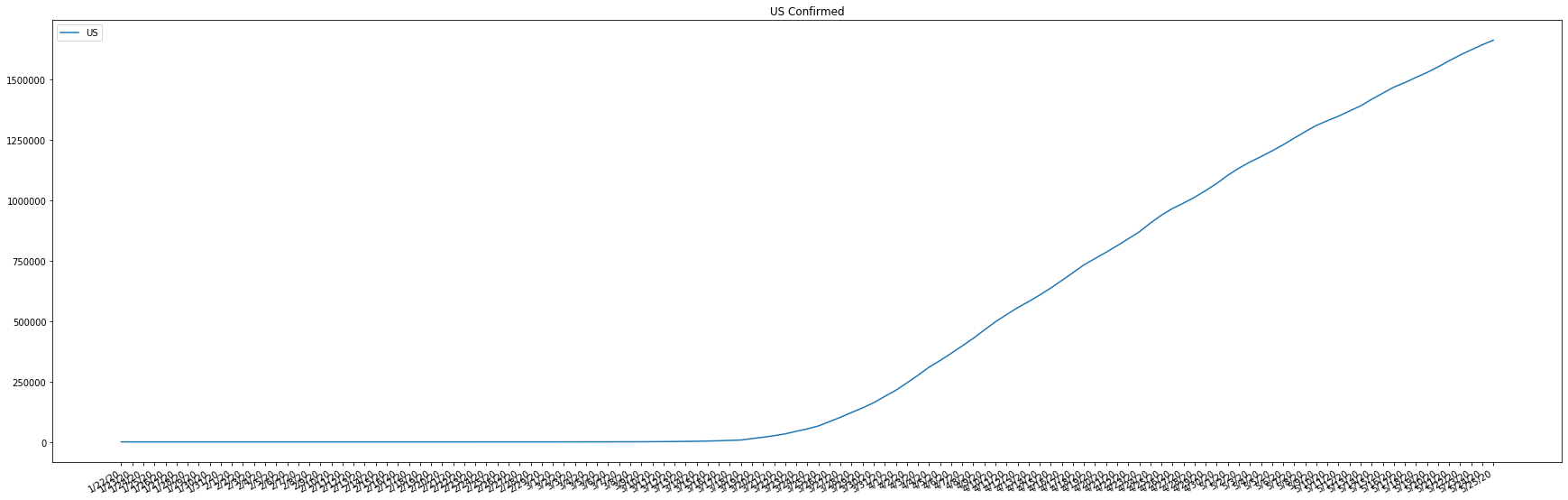
對目前病人數作圖:



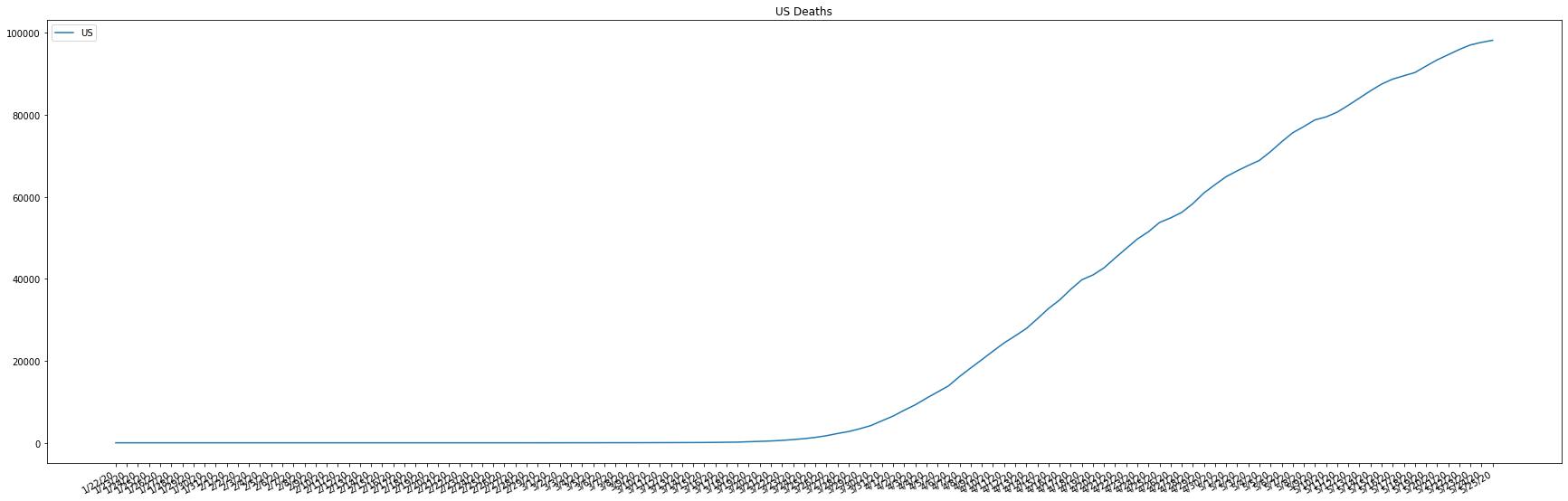
可以看到大部分的趨勢都是一路增加

**＊因為美國人數最多，單獨拿出來看**

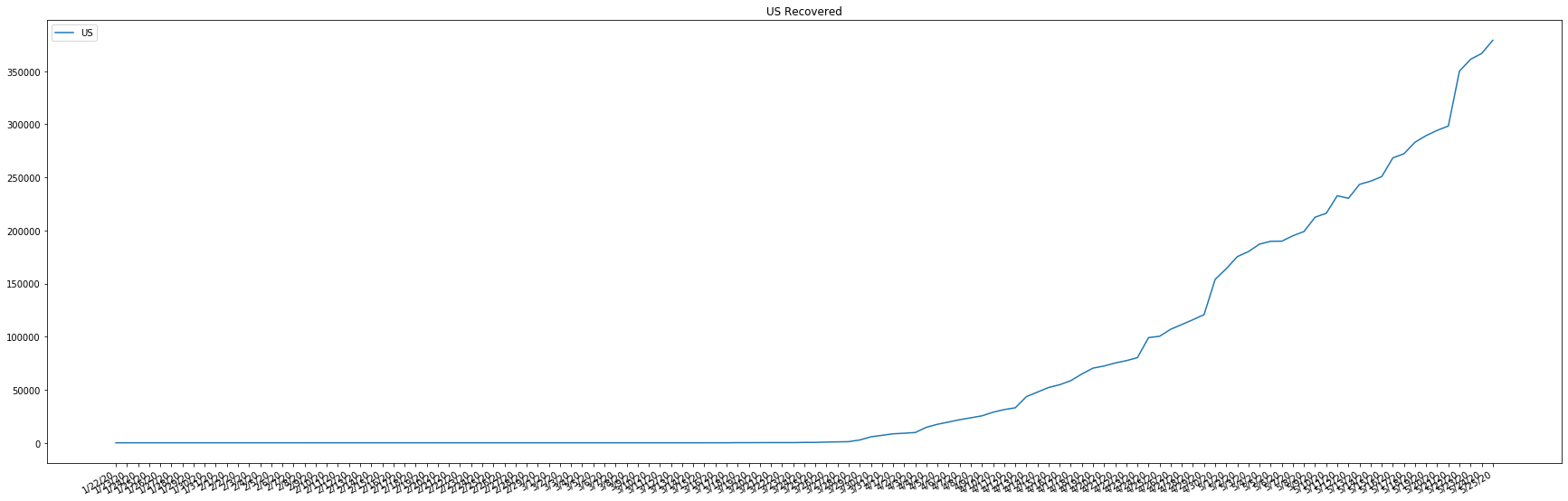
確診:



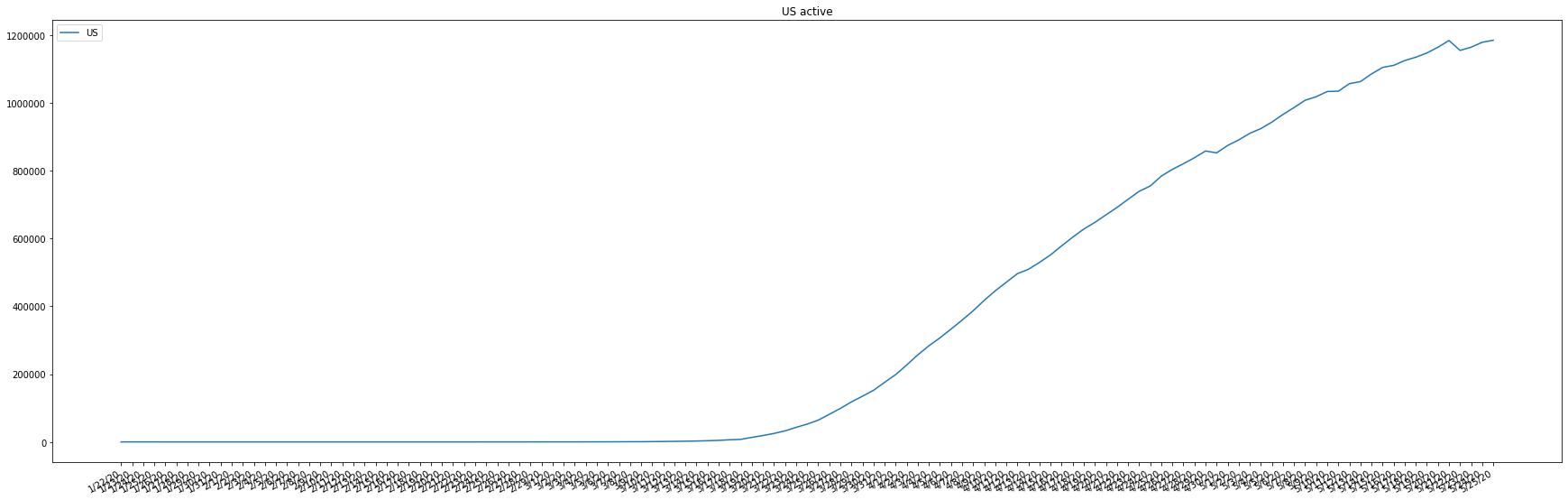
死亡:



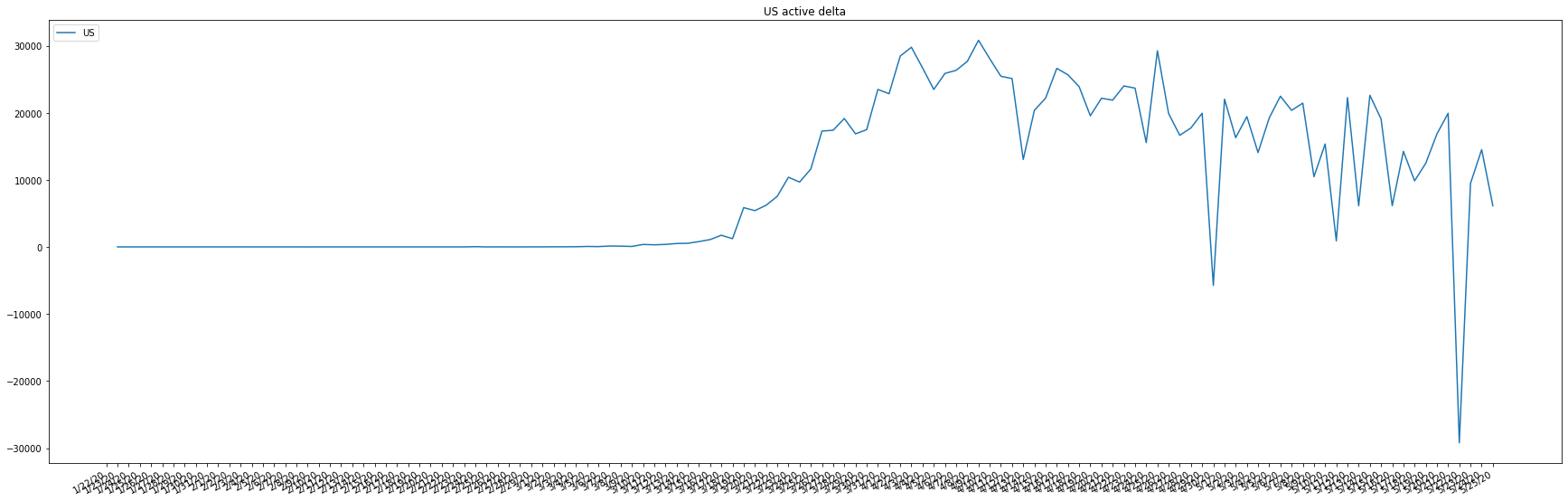
康復:



病人:



ActiveDelta(病人變化)



## 可以發現Activedelta相對來說比較有變化，不像其他都是一路上升，之後拿這種欄位來做預測目標比較有發揮的價值

## 依地理位置作圖:

依日期繪製全球分布動畫

## 一張含有 螢幕擷取畫面 的圖片 自動產生的描述

感染人數變化:

一張含有 畫畫 的圖片

自動產生的描述

死亡人數變化:

一張含有 地圖, 文字 的圖片

自動產生的描述

康復人數變化:

一張含有 文字, 地圖 的圖片

自動產生的描述

病人數變化:

一張含有 文字, 地圖 的圖片

自動產生的描述

## 依國家排名作圖:

## 一張含有 螢幕擷取畫面 的圖片 自動產生的描述

## 感染人數排名:

## 一張含有 螢幕擷取畫面 的圖片 自動產生的描述

## 死亡人數排名:

## 一張含有 螢幕擷取畫面 的圖片 自動產生的描述

## 康復人數排名:

## 一張含有 螢幕擷取畫面 的圖片 自動產生的描述

## 病人數排名:

## 一張含有 螢幕擷取畫面 的圖片 自動產生的描述

1. tests.csv

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

欄位有Country 、 TotalCases、NewCases 、 TotalDeaths 、 NewDeaths 、TotalRecovered 、 ActiveCases 、 Serious,Critical 、 Cases per 1M pop 、Deaths per 1M pop 、 Total Tests 、 Tests per 1M pop 、 Population。

一張含有 桌 的圖片

自動產生的描述

可以發現有大量缺值 -> 填補值:

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

## 依1M pop作圖

## Case per 1M population(一部分)

一張含有 建築物 的圖片

自動產生的描述

## Deaths per 1M population(一部分)

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

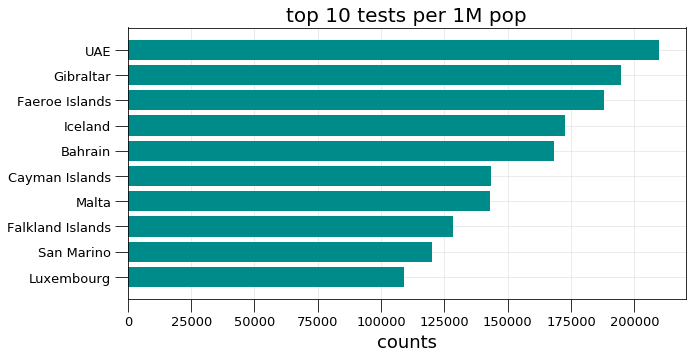
## 依檢測人數作圖

一張含有 螢幕擷取畫面, 白色, 畫畫 的圖片

自動產生的描述

可以看到，usa總共檢測了最多人

但是在平均百萬人受檢數中，USA反而排名連前10都沒有



1. usa\_county\_wise.csv

一張含有 螢幕擷取畫面, 電腦, 螢幕, 差異 的圖片

自動產生的描述

看前三列資料，欄位有UID 、 iso2 、 iso3 、 code3 、 FIPS 、 Admin2 、 Province\_State 、 Country\_Region 、 Lat 、 Long\_ 、 Combined\_Key 、 Date 、 Confirmed 、 Deaths 。

一張含有 螢幕擷取畫面 的圖片

自動產生的描述看各欄位，只有FIPS 、 Admin2有少數空格

列出所有州:

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

## 依地理位置作圖

依日期分類:

一張含有 文字, 地圖 的圖片

自動產生的描述

**一張含有 文字, 地圖 的圖片

自動產生的描述**

1. **Problem define:**

由以上分析過程，可以發現多數數據都是隨時間增加的，如果是要預測的話意義不大，但是activedelta或是各種比率變化才有較大的起伏變化，故暫定以預測activedelta或死亡率的數值優先

## data selection

## 選擇單一國家的為準，這樣標準或特性比較一致，又因美國的確診數最多，故選擇用1.中的美國資料來訓練

## 一張含有 螢幕擷取畫面 的圖片 自動產生的描述

## 一張含有 鳥 的圖片 自動產生的描述

## 2.regression(sklearn)預測死亡率

此問題目標用傳統機器學習方法，解決計算死亡率和額外欄位(前1天死亡數)

## 死亡率:

## 一張含有 螢幕擷取畫面 的圖片 自動產生的描述

## 一張含有 男人, 風箏, 水, 桌 的圖片 自動產生的描述

## 計算相關係數:

## 一張含有 螢幕擷取畫面 的圖片 自動產生的描述

## 選擇特徵Confirmed, preDeaths, Active, Recovered

## 因各項特徵的值都很大故做標準化，並7:3洗牌切割訓練與驗證集

## 一張含有 螢幕擷取畫面 的圖片 自動產生的描述

## 使用LinearRegression，並尋找表現最好的多項式order

## 一張含有 螢幕擷取畫面 的圖片 自動產生的描述

## 結果:

## 

## 評分方式由上而下是r2\_score，MSE，MAE

## 可看出order2時結果最佳

## 繪製比較圖

## 一張含有 風箏, 水, 飛行, 男人 的圖片 自動產生的描述

以上可以看到線性回歸的效果，有一段尖峰沒辦法預測到

所以再改用非線性回歸的**RandomForestRegressor**看看

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

結果:

一張含有 鳥 的圖片

自動產生的描述

比較圖:

一張含有 風箏, 飛行, 水, 彩色 的圖片

自動產生的描述

以上可見表現很好，與train和test資料比較的結果也十分接近，但也可能是資料的樣本太少導致overfitting，故以下增加限制(max\_depth=3, max\_leaf\_nodes=5)減少overfitting

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

結果:

一張含有 室內, 桌 的圖片

自動產生的描述

一張含有 水, 風箏, 船, 彩色 的圖片

自動產生的描述

限制後的r2 score約在0.95

## 3.neural network(LSTM) 以time series 預測Activedelta

## 此問題目標用neural network學習方法解決

## 一張含有 螢幕擷取畫面, 桌 的圖片 自動產生的描述

## 先標準化，在此只取Activedelta作為輸入特徵

## 一張含有 螢幕擷取畫面, 手機 的圖片 自動產生的描述

## 由sequence\_length取平移窗並打亂順序，代表是用前sequence\_length-1個資料來預測最後一個

## 一張含有 螢幕擷取畫面 的圖片 自動產生的描述

## 建立並訓練一個兩層的lstm模型，第一層是多對多lstm，第二層是多對一lstm並加入Dropout防止overfitting，最後輸出一個數值 loss用mse，optimizer用回歸問題常用的rmsprop

## 一張含有 螢幕擷取畫面 的圖片 自動產生的描述

## 繪製實際和預測走勢:

## 一張含有 螢幕擷取畫面 的圖片 自動產生的描述

由上可以發現到大致的走勢有符合

以下換個國家(南韓)的資料再測試

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

一張含有 水, 男人, 房間 的圖片

自動產生的描述

預測結果:

一張含有 文字 的圖片

自動產生的描述

由上可以發現到大致的走勢也有符合，但是這種預測為input的往後shifting的預測結果常常代表了模型的overfitting，可能的原因是資料數太少或模型太複雜

故下面對模型做出一些調整

但是如何sequence\_length的適當長度呢？ 以下也分別用不同長度的sequence\_length來比較預測結果

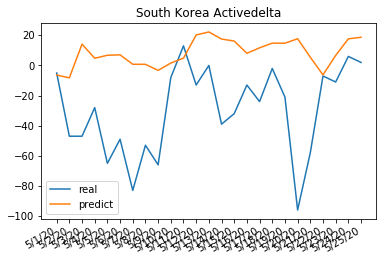
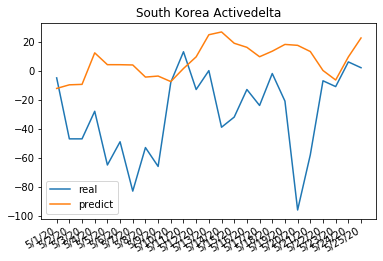
sequence\_length=6 sequence\_length=5

一張含有 文字, 地圖 的圖片

自動產生的描述一張含有 文字 的圖片

自動產生的描述

sequence\_length=4 sequence\_length=3

  
可以發現sequnce\_length=4, 5(代表用前3, 4天來預測下一天)時shifting的現象消失了且也有抓到部分的趨勢，照圖形看起來是sequnce\_length=4時預測效果最好