0 ) 我這次的主題是用環境變數 像是溫度、濕度、雨量、pm2.5數值 對登革熱確診病例的影響

2 ) 我的研究背景和動機就是我再找資料的時候有一天看到電視上寫說登革熱疫情新增了不少病例，而且還有本土群聚，喔好可怕，所以就想說來探討一下到底是甚麼原因造成登革熱，想像中登革熱就是再潮濕、炎熱的環境中會造成的，因此就以環境當成這裡的解釋變數，登革熱確診人數當成我的反應變數。

3 ) 這裡的研究目的就是分析解釋變數&反應變數的基本統計量、相關性等等。最後再進入模型並且評估他的配飾度等等 並且盡可能的修正模型。

所以我們就先決定主題後，變數真的太多了，選取幾個有興趣的出來，並且合併資料，後續再觀察變數，把它丟進檢定、模型裡面看看結果怎麼樣。

4 ) 這裡定義研究的對象，從此圖以及以往的觀念高雄的登革熱會比較嚴重，再加上他是潮溼高溫的氣候，有利於蚊蟲的生長，另外如果是講境外移入變因就有點太多了，所以這裡就鎖定高雄2023整年資料當研究的對象。

5 ) 這裡我們用到疾管屬的登革熱確診數這裡GROUP BY 高雄地區的各行政區像是左營區、仁武區和登革熱確診發病日、高雄地區各行政區人口數這裡的資料粒度是月，但直接把他JOIN到上面的資料中

6 ) 各空氣品質資料一開始她長得像上面這樣對就是有一個COLUM他是全部都是測項、一個小時一筆資料，但這跟我想像中的樣子不一樣，我只需要一天一筆資料然後希望COLUMN是所有的測項，因此我就對每一個測項每一天的資料取最大、最小、平均數，雨量的部分算加總，來得到我們的資料如下。

7 ) 接下來我們就來觀一下我們的反應變數以及解釋變數：有11個連續變數、2個類別變數。

8 ) 我們先從反應變數開始，這裡因為人口總數很大，登革熱確診數很小，因此為了方便觀察我們把單位訂為PPM，才不會每個數值都是0.000000幾這樣，這裡他一定是有確診數才會有資料，所以不會有0的資料，我們這裡可以看得出來她不符合常態。

9 )後續跑了他的分布圖，可以看出100以上的資料不多，因此這裡把它當成離群直，刪除之後也可以比叫看得出來資料的分布

10 ) 然後我們這裡有兩個類別變數、一個是地區，分為南高雄北高雄，以高雄火車站為界線，不是我定義的喔，是真的有這樣區分的。

11 ) 這裡我們分組來看一下散步圖，可以看出來他分布沒有很明顯的區隔，因此這裡我們就沒有分組進行建模

12 ) 第二個類別變數是季節，依照常理把它分為春夏秋冬，這裡我們從統計圖中可以看到大多的資料都集中在秋季，猜測是夏季的時候盛行，然後慢慢傳染秋季病例到達最高峰，冬季就逐漸趨緩，可能因為氣候條件不利於蚊蟲活動，春季沒有登革熱確診病例，當然這裡也不排斥是沒有記錄到資料。

13 ) 這裡我們也來看看他的分組散布圖看看，看起來也沒有因為分組有太大的區隔，所以就也不分組建模了。

14 ) 這裡我們觀察我們十一個解釋變數可以看出綠色底圖的是比較接近常態的，後續可以挑選這些當成主要的解釋變數，紅色的是比較偏離紅線的，可能可以盡量避免選擇這些變數。

15 ) 從這張圖可以看的出來也是綠色的比較接近常態，其中當日最高溫度看起來最符合常態

16 ) 但把他帶入單變量分析跑出來的結果仍然是跟想像中一樣都不符合常態的。

17 ) 第二節我把他帶入假設檢定，我們先從卡方檢定開始，這裡他的虛無假設是這樣的最後我們得到的p value大於 0.05 不拒絕虛無假設，我們沒有足夠證據證明他們有關聯

18 ) 再來我們把只有兩個label的變數”地區”帶入two sample t test 獨立樣本t 檢定，我們可以看到針對F檢定的結果p value 小於 0.05 拒絕虛無假設說明兩個變異數不相等，因此後續的T檢定就看下面的那個不相等的部分，可以看到p value 小於 0.05 拒絕虛無假設 因此我們這裡判定兩組在統計上存在顯著差異。

19 ) 再來我們把它拿去畫圖，可以看出來在南高雄的箱子比較大，如同前面的結果它們是有顯著差異的。

20 ) 接下來把季節帶入anova去作檢定，可以看出p value小於0.05 拒絕虛無假設，在統計上組別之間至少有一組的存在顯著差異，後續也可以檢查每兩組的是否存在顯著差異。

21 ) 因此我們也畫了箱型圖，可以看出來他們彼此之間變異數都有顯著差異，也可以帶入檢定作進一步的確認，對未來，不是現在。

22 ) 這裡我們把他帶入皮爾森相關係數檢定，首先我們可以看出有些相關係數是比較高的，後續可以選取這些變數作為主要變數。另外，可以看到部分變數像是當日最高相對濕度、當日平均溫度 p value 大於 0.05 不拒絕虛無假設，在統計上這裡沒有足夠證據證明他們顯著偏離0.5，其餘的變數在這裡p value都小於0.05 拒絕虛無假設，說明在統計上相關性顯著偏離0.5。

23) 從這張圖可以看出綠色背景的看起來相關性較高，前面也可以看到大約落在0.57~0.58接近高度相關，因此後面選擇這兩個變數作為主要的解釋變數。

24 ) 這裡我們從上面的配飾圖可以看出那條線很曲折，波動很大，想像中可以刪除離群值再去做一次，但現在目前還是再探索性的分析，他又長得太醜了，所以我決定直接查看他非線性關係的分析。從下面可以看出來隨著當日最高相對濕度增加，殘差的變異也逐漸地變大。

25 ) 從上面那張圖相較剛剛的圖形已經是比較趨於平滑的了，但還是可以看得出來解釋變數和反應變數都還是存在一些離群直， y大多數直都是小於60、x大多都是小於35的從想像中台灣也很少會有溫度大於35度的時候，因此這裡決定把他們刪除。可以再重新看到下圖是再刪除過後他有更平滑一點，因此後續再把它帶入線性回歸裡面看看。

26 ) 從這裡回歸模型的結果可以看得出來他的模型是具有統計上的顯著性，而且這裡配飾不足

27 ) 另外我們也可以得到回歸方程式，但這裡的R平方很小，因此後面會針對反應變數進行轉換。從常態性檢定和常數變異數檢定可以看出拒絕他是常態的，且兩群的變異數是不一樣的，且它們是有限性關係的。後續會針對Box-cox結果對 Y 取 ，希望可以改善模型的配適度。

30 ) 接下來我們快速來看結論，轉換後的R平方有提升一丟丟，並且可以從qq plot可以看出來兩端的偏差有縮小，更接近常態一點。但就結果來說R平方還是很差，模型仍存在較大的未解釋變異，所以接下來就是建議的部分。

31 ) 修改變數的分組，目前是高雄各行政區做分組，後面可以用縣市作為分組，這樣環境變數差異也會比較大，雖然可能會犧牲一些細節，但可能更能表示整體趨勢。或是加入更多變數，像是日照時數、都市化指數、家庭收入、教育水平等。都有機會會影響蚊蟲的孳生以及登革熱確診的機率。