Name: Han Hong Tuck from EP0302_01

Title of Data Analysis: The number of dengue cases in Singapore has been brought under control from 2015 to 2018

Url of Dataset used: [https://data.gov.sg/dataset/weekly-number-of-dengue-and-dengue-haemorrhagic-fever-cases (https://data.gov.sg/dataset/weekly-number-of-dengue-and-dengue-haemorrhagic-fever-cases)](https://data.gov.sg/dataset/weekly-number-of-dengue-and-dengue-haemorrhagic-fever-cases)

Questions to answer to gain deeper insights into the chosen datasets

Question 1: Is there an increasing or decreasing trend in the number of dengue cases from 2015 to 2018?

Question 2: Are all the data available/present for the number of dengue cases from 2015 to 2018?

Question 3: How many data points should we plot to show a consistent trend for the number of dengue cases from 2015 to 2018?? / In other words, from which year to which year should we extract the data out of the dataset and plot to display the trend?

Write Python code that uses the Pandas package to extract useful statistical or summary information about the data

In [1]:

```
import pandas as pd

df_dengue_and_dhf_cases = pd.read_csv('weekly-number-of-dengue-and-dengue-haemorr

#to get the first five sets of the pandas dataframe
print(f"First Five Sets of dataset: \n {df_dengue_and_dhf_cases.head(n=10)} \n\n'

#to get the last five sets of the pandas dataframe
print(f"Last Five Sets of dataset: \n{df_dengue_and_dhf_cases.tail(n=10)} \n\n")

#to get details/info about the pandas dataframe
print(f"\n Dataframe Info: \n{df_dengue_and_dhf_cases.info(verbose=bool)}\n")

#to get info on the number of rows and columns about the pandas dataframe
print(f"\n Number of rows and columns: \n{df_dengue_and_dhf_cases.shape}\n")

#to get summary statistics for active practice and non-active practice chinese me
df_dengue_and_dhf_cases_stats = df_dengue_and_dhf_cases.groupby(["type_dengue"])
print(f"Summary Statistics for dengue-and-dengue-haemorrhagic-fever-cases indivi

#to get summary statistics for active practice and non-active practice chinese me
df_yearly_dengue_and_dhf_cases_stats = df_dengue_and_dhf_cases.groupby(["year","t
print(f"Summary Statistics for the number of dengue and dhf cases every year: \n\
```

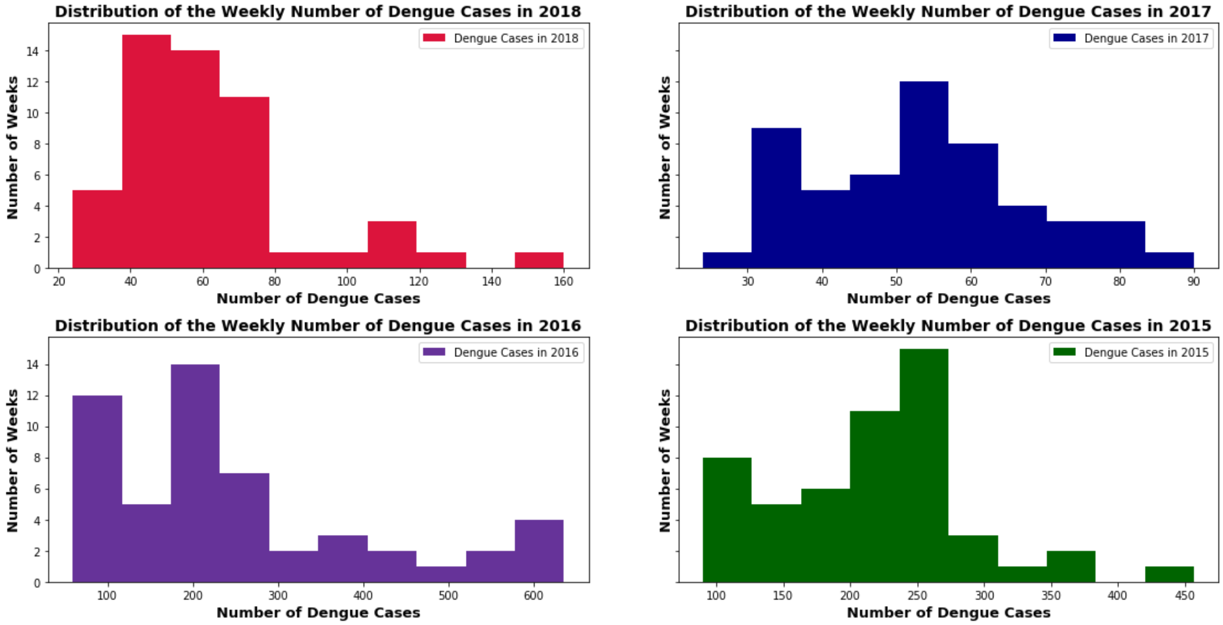
```
First Five Sets of dataset: 
       eweek type_dengue  number
year                            
2014      1      Dengue   436.0
2014      1         DHF     1.0
2014      2      Dengue   479.0
2014      2         DHF     0.0
2014      3      Dengue   401.0
2014      3         DHF     0.0
2014      4      Dengue   336.0
2014      4         DHF     0.0
2014      5      Dengue   234.0
2014      5         DHF     0.0 


Last Five Sets of dataset: 
      eweek type_dengue  number
year                           
2018     49      Dengue   113.0
2018     49         DHF     1.0
2018     50      Dengue   107.0
2018     50         DHF     1.0
2018     51      Dengue   127.0
2018     51         DHF     1.0
2018     52      Dengue   160.0
2018     52         DHF     0.0
2018     53      Dengue     NaN
2018     53         DHF     NaN
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 530 entries, 2014 to 2018
Data columns (total 3 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   eweek        530 non-null    int64  
 1   type_dengue  530 non-null    object 
 2   number       522 non-null    float64
dtypes: float64(1), int64(1), object(1)
memory usage: 16.6+ KB

 Dataframe Info:
None


 Number of rows and columns:
(530, 3)

Summary Statistics for dengue-and-dengue-haemorrhagic-fever-cases individually:

type_dengue          DHF      Dengue
number count  261.000000  261.000000
       mean     0.379310  186.421456
       std      0.654712  158.706595
       min      0.000000   24.000000
       25%      0.000000   60.000000
       50%      0.000000  157.000000
       75%      1.000000  250.000000
       max      4.000000  888.000000


Summary Statistics for the number of dengue and dhf cases every year:

year                2014                   2015                   2016  \
type_dengue          DHF      Dengue        DHF      Dengue        DHF   
number count  53.000000   53.000000  52.000000   52.000000  52.000000   
       mean    0.377358  345.396226   0.230769  216.961538   0.461538   
       std     0.627155  174.097822   0.469267   75.448193   0.778675   
       min     0.000000  149.000000   0.000000   90.000000   0.000000   
       25%     0.000000  212.000000   0.000000  168.250000   0.000000   
       50%     0.000000  291.000000   0.000000  225.500000   0.000000   
       75%     1.000000  436.000000   0.000000  256.250000   1.000000   
       max     3.000000  888.000000   2.000000  457.000000   4.000000   

year                             2017                 2018             
type_dengue       Dengue          DHF     Dengue        DHF      Dengue  
number count   52.000000  52.000000  52.000000  52.000000   52.000000  
       mean   251.173077   0.326923  52.884615   0.500000   62.634615  
       std    160.137897   0.550264  14.649706   0.779643   25.638535  
       min     59.000000   0.000000  24.000000   0.000000   24.000000  
       25%    132.250000   0.000000  40.000000   0.000000   47.250000  
       50%    217.000000   0.000000  51.000000   0.000000   56.000000  
       75%    302.750000   1.000000  62.250000   1.000000   74.250000  
       max    636.000000   2.000000  90.000000   3.000000  160.000000  
```

Write Python code that uses Matplotlib package to produce useful data visualizations that explain the data.

In [2]:

```
import pandas as pd
import matplotlib.pyplot as plt

years_ = [2018,2017,2016,2015]

df = pd.read_csv('weekly-number-of-dengue-and-dengue-haemorrhagic-fever-cases.csv

#using the & operator to check whether it specifies both conditions
#First condition - type of case is dengue
#Second condition - dengue cases that exists from range 2015 to 2018

df = df[(df.type_dengue=="Dengue") & (df.year.isin(years_))].dropna()

colors = ["crimson","darkblue","rebeccapurple","darkgreen"]

fig,ax = plt.subplots(2,2,figsize=(16,8),sharey=True)

#to collapse the array of subplots from two dimension into one dimension
ax = ax.flatten()

#using loops to iterate over the number of dengue cases every year and display th
for i in range(len(years_)):
    ax[i].hist(df[df.year==years_[i]].number,label="Dengue Cases in "+str(years_[
    ax[i].set_xlabel("Number of Dengue Cases",fontsize=13,fontweight="bold")
    ax[i].set_ylabel("Number of Weeks",fontsize=13,fontweight="bold")
    ax[i].set_title("Distribution of the Weekly Number of Dengue Cases in "+str(y
    ax[i].legend()

#adjust spacing between subplots to minimize the overlaps.
fig.tight_layout()

plt.show()
```



For each dataset, explain the nature of that dataset (i.e. what is in that dataset) or any pecularities about it you wish to highlight and explain the process you went through to analyse that dataset, . Where possible, you should specifically mention how you used the Pandas or Matplotlib functions to achieve a certain outcome e.g. to transform the data or to produce a certain visualization:

Pecularities to highlight:

One peeularity to highlight is even though there are many dengue cases every week, the number of dengue and haemorrhagic fever cases has been relatively small. The highest number of dhf in a week is 4 cases compared to 888 cases of dengue cases.

Nature of dataset:

The nature of the dataset consists of the number of dengue cases and dengue-haemorrhagic fever cases every week from 2014 to 2018. After using the .describe() method from pandas to retrieve information about the total dengue and dhf cases individually, I am able to tell that the number of dhf cases in singapore from 2014 to 2018 is relatively low with only a maximum case of 4 dhf cases in a week and almost every week, there are no to 1 or 2 dhf cases. In contrast, there are many dengue cases in singapore every week with a mean average case of arounfd 180 cases. The highest number of dengue cases in Singapore that happened in a week is at 888 cases. By looking at the number of cases individually, I decided to plot the dengue cases as the number of dhf cases is too insignificant to plot on the same histogram with the number of dengue cases. Upon doing so, I have to consider what is the best way to compare the results from the histogram. After I have tried to display the dengue cases of the different years on the same x-axis, I find that there is a very huge spread of data which could lead to inaccuracies when we are using graph to analyse for information. As such, I have decided to use small multiples so that I am able to show the histogram of every year so that I am are able to compare the number of dengue cases from 2015 to 2018 side by side. This will lead to better data analysis as we are able to compare the number of weeks of which the dengue cases exceeds a certain amount (for example the number of dengue cases that exceeds 80 in 2018 subplot comapared with 2017).

Process of using Pandas or Matplotlib functions to transform the data:

The dataset consists of the columns: week, type_year and the number of cases. In order to retrieve the data that is type dengue and from 2015 to 2018, I use the binary "&" operator and specify both conditions for type as dengue and year as from 2015 to 2018 and drop values that are NaN. Then, I specified the color for each subplots, and declare a figure and axes objects with two rows and two columns such that there will be 4 smaller subplots, each showing different data of the number of dengue cases every year. Then, I declare sharey= True so that all the y-axis will show the same range yticks. After doing so, I collapse the array of subplots from two dimension into one dimension so that when I iterate over the dengue cases from 2015 to 2018, the value "i" will be used to plot the histograms. Similarly, I set the title, xlabel, ylabel and the legend during the loop. After doing so, I call the method tight_layout() to adjust the spacing between subplots to minimize the overlaps.

For each dataset, highlight the insights you have gained from analysing the data and any conclusions or recommendations you want to make as a result of the analysis:

After plotting the graph, I can tell that the number of dengue cases that happen every year has been brought down to a significant amount but number of dengue cases that happen every year is still relatively inconsistent. If we only look at the most number of weeks at which the dengue cases are hovering at every year, the number of dengue cases has reduced by 50 cases from 250 in 2015 to 2016. From 2016 to 2017, it decreases even more significantly from 200 cases to arounf 50-60 cases but then rise again in 2018 between 40 to 80 cases. As such, I am able to conclude that there is some progress achieved in bringing down the number of dengue cases but the number of dengue cases seems to spike back in 2018. In order to answer the title of the data analysis, I suggest that the dataset also include the year 2019 for comparison so that I am able to tell whether the number of dengue cases has been brought down or is going to increase again.