Name: Han Hong Tuck from EP0302_01

Title of Data Analysis: When the number of traditional chinese active medical practitioners increase, the number of traditional chinese non-active medical practitioners increases.

Url of Dataset used: https://data.gov.sg/dataset/number-of-traditional-chinese-medicine-practitioners?view_id=a13fefc5-ba15-46b6-8f2c-e693ca9c73ff&resource_id=94ba6f5e-c319-4628-b66d-3ad64a91443c (https://data.gov.sg/dataset/number-of-traditional-chinese-medicine-practitioners?view_id=a13fefc5-ba15-46b6-8f2c-e693ca9c73ff&resource_id=94ba6f5e-c319-4628-b66d-3ad64a91443c)

Questions to answer to gain deeper insights into the chosen datasets

Question 1: What is the relationship between the number of active and non active traditional chinese medical practitioners?

Question 2: Are all the data available/present for the active and non active traditional chinese medical practitioners from 2006 to 2019?

Question 3: How much data should we plot on the graph to show a consistent trend between the number of active and non-active traditional chinese medical practitioners?

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_chinese_pract = pd.read_csv('number-of-traditional-chinese-medicine-practitior

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

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

#to get details/info about the pandas dataframe
print(f"\n Dataframe Info: \n{df_chinese_pract.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_chinese_pract.shape}\n\n")

#to get summary statistics for all data
print(f"\n Summary Statistics for all data: \n\n{df_chinese_pract.describe()}\n\r

#to get summary statistics for active practice and non-active practice chinese me
df_chinese_pract_stats = df_chinese_pract.groupby(["sector"])[["count"]].describe
print(f"Summary Statistics for active-practice and non-active practice chinese me
```

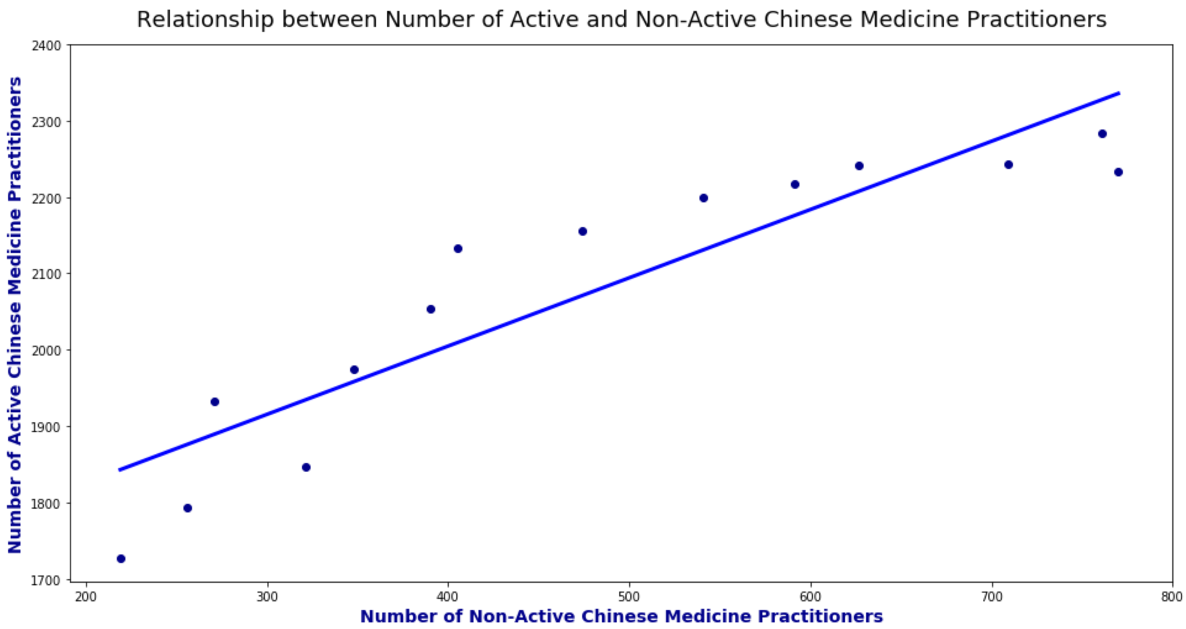
```
First Five sets of dataset: 
                       sector  count
year                                
2006           Active Practice   1727
2006    Not in Active Practice    219
2007           Active Practice   1794
2007    Not in Active Practice    256
2008           Active Practice   1846
2008    Not in Active Practice    321
2009           Active Practice   1932
2009    Not in Active Practice    271
2010           Active Practice   1974
2010    Not in Active Practice    348 


Last Five sets of dataset: 
                      sector  count
year                               
2015           Active Practice   2217
2015    Not in Active Practice    591
2016           Active Practice   2241
2016    Not in Active Practice    627
2017           Active Practice   2243
2017    Not in Active Practice    709
2018           Active Practice   2234
2018    Not in Active Practice    770
2019           Active Practice   2284
2019    Not in Active Practice    761 


<class 'pandas.core.frame.DataFrame'>
Int64Index: 28 entries, 2006 to 2019
```

```
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   sector  28 non-null     object
 1   count   28 non-null     int64 
dtypes: int64(1), object(1)
memory usage: 672.0+ bytes

 Dataframe Info:
None


 Number of rows and columns:
(28, 2)



 Summary Statistics for all data:

             count
count    28.000000
mean   1275.571429
std     833.671284
min     219.000000
25%     456.750000
50%    1248.500000
75%    2138.500000
max    2284.000000


Summary Statistics for active-practice and non-active practice chinese medica
l practitioners individually:

sector       Active Practice  Not in Active Practice
count count        14.000000               14.000000
      mean       2073.785714              477.357143
      std         186.799431              190.444665
      min        1727.000000              219.000000
      25%        1942.500000              327.750000
      50%        2144.000000              439.500000
      75%        2229.750000              618.000000
      max        2284.000000              770.000000
```

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

In [2]:

```
import pandas as pd
import numpy as np
from numpy.polynomial.polynomial import polyfit
import matplotlib.pyplot as plt

#read from file to get dataset
df_chinese_pract = pd.read_csv('number-of-traditional-chinese-medicine-practition

#get data only for active and non_active chinese medical pract from dataset
df_active, df_not_active = df_chinese_pract[df_chinese_pract.sector=="Active Prac

#declare fig and ax object for plotting
fig, ax = plt.subplots(figsize=(16,8))

#plot points for non_active and active medical pract using scatter method
ax.scatter(df_not_active["count"],df_active["count"],color="darkblue")

#plotting best fit line from dataset

#convert medical pract from series to numpy array
np_active,np_not_active = df_active["count"].to_numpy(), df_not_active["count"].t

#plotting the best fit line
#using np.unique to handle the case whereby the x values isn't sorted
#using poly1d to return a function for the line of best fit, which you then evalu
ax.plot(np.unique(np_not_active), np.poly1d(np.polyfit(np_not_active, np_active, 

ax.set_xlabel('Number of Non-Active Chinese Medicine Practitioners',fontweight="b

ax.set_ylabel('Number of Active Chinese Medicine Practitioners',fontweight="bold"

ax.set_title("Relationship between Number of Active and Non-Active Chinese Medici

ax.set_xticks([200,300,400,500,600,700,800]), ax.set_yticks([1700,1800,1900,2000,

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:

Nature of dataset:

The nature of the dataset consists of the number of active and non active traditional chinese medical practitioners from 2006 to 2019. After analysing the dataset using .head() and .tail() method, I am able to tell that there is a general increase in the number of active and non active traditional chinese medical practitioners over the period of time. Using the .info() method, I am also able to tell that all the data are present as there are no null values. In order to find the correlation between active and non active chinese medical pratitioners, I decided to plot a scatter graph to show the relationship.

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

The dataset consists of the columns: year, sector (active and non active traditional chinese medical practitioners) and count (number of medical practitioners). Firstly, I retrieve the data for the active and non_active chinese medical pract from the dataset using the boolean method by specifying the sector that I am trying to retrieve as "Active Practice" and "Not in Active Practice". Then, I declare the figure and axes object to plot points for non_active and active medical pract using scatter method. In order to better represent the relationship between them, I also plotted a best fit line. To do that, I have to convert the pandas series to numpy array where I pass in the (sorted values of x and poly1d which returns a function for the line of best fit from polyfit, mutiplied by the sorted values of x)

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 am able to tell that when the number of active traditional chinese medical practitioners increases, the number of non active traditional practitioners increase. The line of best fit further supports this statement as there is a positive and linear correlation between the number of active chinese medical practitioners and non-active chinese medical practitioners. Hence, we are able to identify that our title of the data analysis shows the correct relationship between active and non-active chinese medical practitioners. However, one limitiation of this dataset is it only shows the data from 2006 to 2019 annually. In order to gather more points to plot on the scatter plot, the dataset also should include data every half a year or quarterly so that within the same timeframe, there will be more points to plot on the scatter plot.