Name: Han Hong Tuck from EP0302_01

Title of Data Analysis: When the gdp percentage increases, the overall spending in the government health expenditure increases as well.

Url of dataset used: https://data.gov.sg/dataset/government-health-expenditure (https://data.gov.sg/dataset/government-health-expenditure)

Questions to answer to gain deeper insights into the chosen datasets

Question 1: Is there an increasing or decreasing trend for the government health expenditure from the year 2006 to 2017?

Question 2: Are all the data available/present for operating expenditure, development expenditure, government health expenditure, as well as percentage gdp?

Question 3: How many data points should we plot to show a consistent trend for the government health expenditure from the year 2006 to 2017?/ 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_gov_expenditure = pd.read_csv('government-health-expenditure.csv',index_col=0)

#to get the first five rows of the pandas dataframe
print(f"First Five Rows of dataset: \n {df_gov_expenditure.head()} \n\n")

#to get the last five rows of the pandas dataframe
print(f"Last Five Rows of dataset: \n{df_gov_expenditure.tail()} \n\n")

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

#to get summary statistics for operating_expenditure,development_expenditure,gove

df_gov_expenditure_stats = df_gov_expenditure.describe()

print(f"Summary Statistics for Government Health Expenditure: \n\n{df_gov_expendi
```

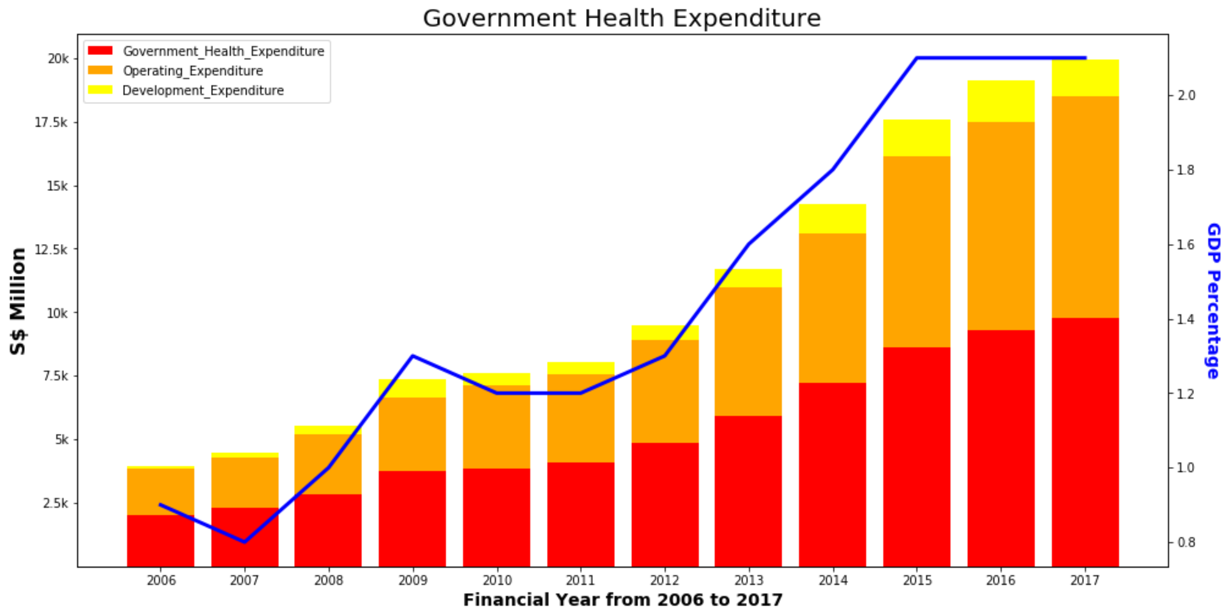
```
First Five Rows of dataset: 
                 operating_expenditure  development_expenditure  \
financial_year                                                    
2006                             1840                       96   
2007                             2019                      185   
2008                             2379                      336   
2009                             2920                      711   
2010                             3258                      485   

                government_health_expenditure  percentage_gdp  
financial_year                                                 
2006                                   2009.7             0.9  
2007                                   2283.2             0.8  
2008                                   2814.1             1.0  
2009                                   3745.8             1.3  
2010                                   3856.7             1.2   


Last Five Rows of dataset: 
                operating_expenditure  development_expenditure  \
financial_year                                                    
2013                             5044                      723   
2014                             5872                     1147   
2015                             7520                     1413   
2016                             8199                     1618   
2017                             8734                     1465   

                government_health_expenditure  percentage_gdp  
financial_year                                                 
2013                                   5938.1             1.6  
2014                                   7223.1             1.8  
2015                                   8639.9             2.1  
2016                                   9307.0             2.1  
2017                                   9764.3             2.1  
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12 entries, 2006 to 2017
Data columns (total 4 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   operating_expenditure           12 non-null     int64  
 1   development_expenditure         12 non-null     int64  
 2   government_health_expenditure   12 non-null     float64
 3   percentage_gdp                  12 non-null     float64
dtypes: float64(2), int64(2)
memory usage: 480.0 bytes

 Dataframe Info:
None


 Number of rows and columns:
(12, 4)


Summary Statistics for Government Health Expenditure:

       operating_expenditure  development_expenditure  \
count              12.000000                12.000000   
mean             4611.666667               769.750000   
std              2444.613497               518.493644   
min              1840.000000                96.000000   
25%              2784.750000               423.750000   
50%              3777.500000               658.000000   
75%              6284.000000              1213.500000   
max              8734.000000              1618.000000   

       government_health_expenditure  percentage_gdp  
count                      12.000000       12.000000  
mean                     5375.891667        1.450000  
std                      2754.457501        0.477684  
min                      2009.700000        0.800000  
25%                      3512.875000        1.150000  
50%                      4464.400000        1.300000  
75%                      7577.300000        1.875000  
max                      9764.300000        2.100000  
```

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

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.ticker as ticker

df_gov_exp = pd.read_csv('government-health-expenditure.csv',index_col=0)

#declare the figure and axes object to plot
fig, ax = plt.subplots(figsize=(16,8))

#change the xticks frequency on the x-axis
ax.xaxis.set_major_locator(ticker.MultipleLocator(1))

#replace the yticks on the y-axis
plt.yticks([2500, 5000, 7500, 10000, 12500, 15000, 17500, 20000],['2.5k', '5k', '

#create variable data that contains a numpy array of government health expenditur
data = np.array([df_gov_exp.government_health_expenditure, df_gov_exp.operating_e

#specify different colors for the bars of government health expenditure, operatin
colors=["red","orange","yellow"]

#create a vertically stacked numpy array in sequcnce with the first array being a
#to the shape of the first array of data which is then followed by government hea
#development expenditure such that when bottom will always return an array of ind
bottom = np.vstack((np.zeros((data.shape[1])),np.cumsum(data, axis=0)))

#specify different labels for the bars of government health expenditure, operatin
labels=["Government_Health_Expenditure","Operating_Expenditure","Development_Expe

#using a loop to iterate over the variables (data,colors,bottom and labels) to pl
for dat, col, bot, lab in zip(data,colors,bottom,labels):
    ax.bar(df_gov_exp.index, dat, color=col, bottom=bot, label=lab)

#Create a twin Axes that shares the x-axis
ax2 = ax.twinx()

#plot the twin axes on the same graph
ax2.plot(df_gov_exp.index,df_gov_exp["percentage_gdp"],color="blue",linewidth=3)

#set the ylabel of the twin axes
ax2.set_ylabel("GDP Percentage",color="blue",rotation=270,labelpad=20,fontsize=14

#to set title and label for x-axis and y-axis on the graph
ax.set_title("Government Health Expenditure",fontsize=20)
ax.set_xlabel("Financial Year from 2006 to 2017",fontsize=14,fontweight="bold"),

#Display the legend
ax.legend()

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 pecularity to highlight is that the trend of the gdp percentage is relatively inconsistent. From 2006 to 2007, the percentage gdp showed a sharp decline but rose back again at a constant rate from 2007 to 2009. This is then followed by a dip in the gdp percentage again from 2009 to 2010. Then, the gdp percentage remains the same for the year 2010 and 2011, and then again showing a very slight increase from year 2010 and 2011. From 2012 to 2015, the increase in the gdp percentage becomes much steeper compared to the growth rate before. However, the percentage gdp reaches a standstill from the year 2015 to 2017, not showing any growth.

Another pecularity to highlight is that the general trend of the gdp percentage does not correspond to the overall trend of the government health expenditure. From the graph, we can tell that the growth in the government health expenditure has been relatively constant, increasing at a constant rate. When we compared this overall trend with the gpd percentage, we are able to tell that there is no signifiant correlation between them such as the increase in the government expenditure from 2006 to 2007 but a sharp decline in gdp percentage. Similarly, the slight dip in the gdp percentage from 2009 to 2010 also has a slight increase in the governmenr health expenditure. Another point to note is there is an increase in the government health expenditure from 2015 to 2017 but yet no increase in the gdp percentage during that period.

Nature of dataset:

The nature of the dataset consists of the amount of government health expenditure (operating expenditure, development expenditure and governemnt health expenditure and gdp percentage from 2006 to 2017). After using pandas to extract the data using .head() and .tail() method. I found out that there is increase in all the four columns over the years (operating expenditure, development expenditure, government health expenditure as well as gdp percentage). Since all three columns show similar growth rate and all data are available in the dataset, I decided to plot a bar chart.

After comparing the advantages of plotting a stacked bar chart and multi-series bar chart, I decided to go with the stacked bar chart because they are easier to read and analyse for data visualisation. If I have plotted a multi-series bar chart, it will be even more harder to read because there will be many too many bar diagrams and it will not be effective in explaining the dataset itself

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

Firsly, I declare a figure and axes object to plot the data and specify xticks frequency to 1 using ticker.MultipleLocator method. Then, I replace the yticks by changing it to display in terms of k(every 1000) so that it will be easier to read. An example would be to replace 2500 to 2.5k when displaying on the yticks. Then I create the variable data that contains a numpy array of government health expenditure, operating expenditure and development expenditure. Afterwards, I specify different colors for the bars of government health expenditure, operating expenditure and development expenditure so that it will be easier to differentiate between the expenditures. Then, I create a vertically stacked numpy array in sequcnce with the first array being a empty list of values identical to the shape of the first array of data. This is then followed by government health expenditure, operating expenditure and development expenditure such that bottom will always return an array of index -1 with reference to the data variable. Then, I specify different labels for the bars of government health expenditure, operating expenditure and development expenditure and used a loop to iterate over the variables (data,colors,bottom and labels) to plot a stacked bar chart.

In addition, I also create a twin Axes that shares the x-axis so that I am able to plot the gdp percentage on the same graph. By doing so, I am able to tell the relationship between government expenditure and the gdp percentage. Finally, I set the ylabel of the twin axes, title and label for x-axis and y-axis and display the legend.

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 learned that generally when the gdp percentage increases, the government expenditure increases as well. However, there are a few anomailies on the graph as there are a few instances where the increase in the government health expenditure has a decreased or unchanged gdp percentage in the corresponding year. One such example is that the amount of government health expenditure increases from 2015 to 2017, but the gdp percentage remains stagnant at 2.1% throughout the three years. One recommendation would be to include more data into the dataset (e.g. from 1980 to 2017) so that when there is more data plotted, we are able to identify a more consistent trend for the relationship between government health expenditure and percentage gdp.