lab 1 assignment export

August 11, 2021

1 Basic Descriptive Analytics with Python

This notbook is a submission for CSDA1010 - Basic Methods of Data Analytics course asignement individual asignement one.

[87]:
```
# import the necessary packages to manipulalate data
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

[88]:
```
#set working directory as the current working directory
os.chdir(os.getcwd())
```

[89]:
```
# import data as a pandas dataframe
# IBMStock.csv, GEStock.csv, ProcterGambleStock.csv, CocaColaStock.csv,␣
 ↪BoeingStock.csv

IBM = pd.read_csv('data/IBMStock.csv')
GE = pd.read_csv('data/GEStock.csv')
ProcterGamble = pd.read_csv('data/ProcterGambleStock.csv')
CocaCola = pd.read_csv('data/CocaColaStock.csv')
Boeing = pd.read_csv('data/BoeingStock.csv')
```

[90]:
```
GE.head()
```

[90]:

	Date	StockPrice
0	1/1/70	74.253333
1	2/1/70	69.976842
2	3/1/70	72.158571
3	4/1/70	74.252727
4	5/1/70	66.665238

1.0.1 Change datatypes for all stock dataframes

- Change Date datatype as a native datetime rather than the default object

[91]:
```
IBM['Date'] = pd.to_datetime(IBM['Date'], format='%m/%d/%y')
GE['Date'] = pd.to_datetime(GE['Date'], format='%m/%d/%y')
ProcterGamble['Date'] = pd.to_datetime(ProcterGamble['Date'], format='%m/%d/%y')
CocaCola['Date'] = pd.to_datetime(CocaCola['Date'], format='%m/%d/%y')
Boeing['Date'] = pd.to_datetime(Boeing['Date'], format='%m/%d/%y')
```

[92]:
```
# Check the data after data type adjstement
IBM.head()
```

[92]:

	Date	StockPrice
0	1970-01-01	360.319048
1	1970-02-01	346.723684
2	1970-03-01	327.345714
3	1970-04-01	319.852727
4	1970-05-01	270.375238

[93]:
```
GE.head()
```

[93]:

	Date	StockPrice
0	1970-01-01	74.253333
1	1970-02-01	69.976842
2	1970-03-01	72.158571
3	1970-04-01	74.252727
4	1970-05-01	66.665238

[94]:
```
ProcterGamble.head()
```

[94]:

	Date	StockPrice
0	1970-01-01	111.874286
1	1970-02-01	111.453684
2	1970-03-01	108.451429
3	1970-04-01	106.288636
4	1970-05-01	73.332857

[95]:
```
CocaCola.head()
```

[95]:

	Date	StockPrice
0	1970-01-01	83.368095
1	1970-02-01	81.591053
2	1970-03-01	81.338095
3	1970-04-01	76.805909
4	1970-05-01	69.278571

[96]:
```
Boeing.head()
```

[96]:

	Date	StockPrice
0	1970-01-01	27.853810
1	1970-02-01	22.381053

```
2 1970-03-01   23.105238
3 1970-04-01   21.571364
4 1970-05-01   18.932857
```

2 Warm-up/Basic statistics Questions

How many rows of data are in each dataset?

[97]:
```
print("There are {} rows and {} columns in IBMStock dataset". format(IBM.
 ↪shape[0], IBM.shape[1]))
print("There are {} rows and {} columns in GEStock dataset". format(GE.
 ↪shape[0], GE.shape[1]))
print("There are {} rows and {} columns in ProcterGambleStock dataset".␣
 ↪format(ProcterGamble.shape[0], ProcterGamble.shape[1]))
print("There are {} rows and {} columns in CocaColaStock dataset".␣
 ↪format(CocaCola.shape[0], CocaCola.shape[1]))
print("There are {} rows and {} columns in BoeingStock dataset". format(Boeing.
 ↪shape[0], Boeing.shape[1]))
```

```
There are 480 rows and 2 columns in IBMStock dataset
There are 480 rows and 2 columns in GEStock dataset
There are 480 rows and 2 columns in ProcterGambleStock dataset
There are 480 rows and 2 columns in CocaColaStock dataset
There are 480 rows and 2 columns in BoeingStock dataset
```

What is the earliest/latest year in our datasets?

[98]:
```
# Finding the earliest and latest year in our dataset
print("The  earliest year is IBMStock dataset {}".format(IBM['Date'].dt.year.
 ↪min()))
print("The  latest year is IBMStock dataset {}".format(IBM['Date'].dt.year.
 ↪max()))
print("\n")
print("The  earliest year is GEStock dataset {}".format(GE['Date'].dt.year.
 ↪min()))
print("The  latest year is GEStock  dataset {}".format(IBM['Date'].dt.year.
 ↪max()))
```

```
The  earliest year is IBMStock dataset 1970
The  latest year is IBMStock dataset 2009


The  earliest year is GEStock dataset 1970
The  latest year is GEStock  dataset 2009
```

For the period above what is the average stock price of Coca Cola?

[99]:
```
# The avaregae stock price for Coca Cola from Year 1970 to 2009
print("The average Coca Cola is {}".format(CocaCola.StockPrice.mean()))
```

```
The average Coca Cola is 60.02972973327079
```

What is the maximum price of IBM during this period?

[100]:
```
print("The maximum price of IBM is {}".format(IBM.StockPrice.max()))
```

```
The maximum price of IBM is 438.9015789
```

What is the standard deviation of P&G stock price over this period?

[101]:
```
print("The standard deviation of P&G stock is {}".format(ProcterGamble.
↪StockPrice.std()))
```

```
The standard deviation of P&G stock is 18.19414030797156
```

What is the median price of Boeing in the last 5 years for which we have data?

[102]:
```
# To get the last five years, subtract 5 from the max year

last_five_year_data = Boeing[Boeing.Date.dt.year > Boeing.Date.dt.year.max() -␣
↪5]

print("The median price of Boeing Stock price in the last 5 years is {}".
↪format(last_five_year_data.StockPrice.median()))
```

```
The median price of Boeing Stock price in the last 5 years is 69.675666665
```

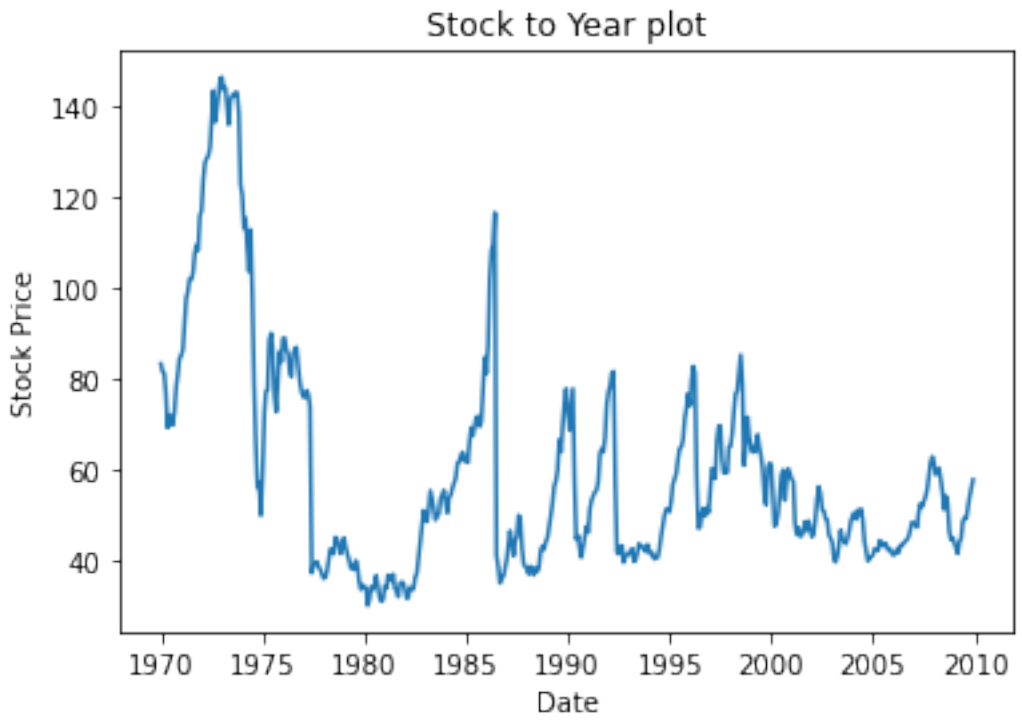
3 Basic Plotting Questions

3.1 Part 1:

Plot the StockPrice of Coca-Cola on the Y-axis across Date on the x-axis using the basic plot() fuction. What do you see when you use the default plot function what do you see? Scatter-plot, eh?

[103]:
```
plt.xlabel('Date')
plt.ylabel("Stock Price")
plt.title("Stock to Year plot")
plt.plot(CocaCola.Date, CocaCola.StockPrice)
```

[103]:
```
[<matplotlib.lines.Line2D at 0x190345c44c0>]
```



Based on the above line plot, we can answer the following two questions

Identify the year during which Coca-Cola had the highest/lowest stock price?

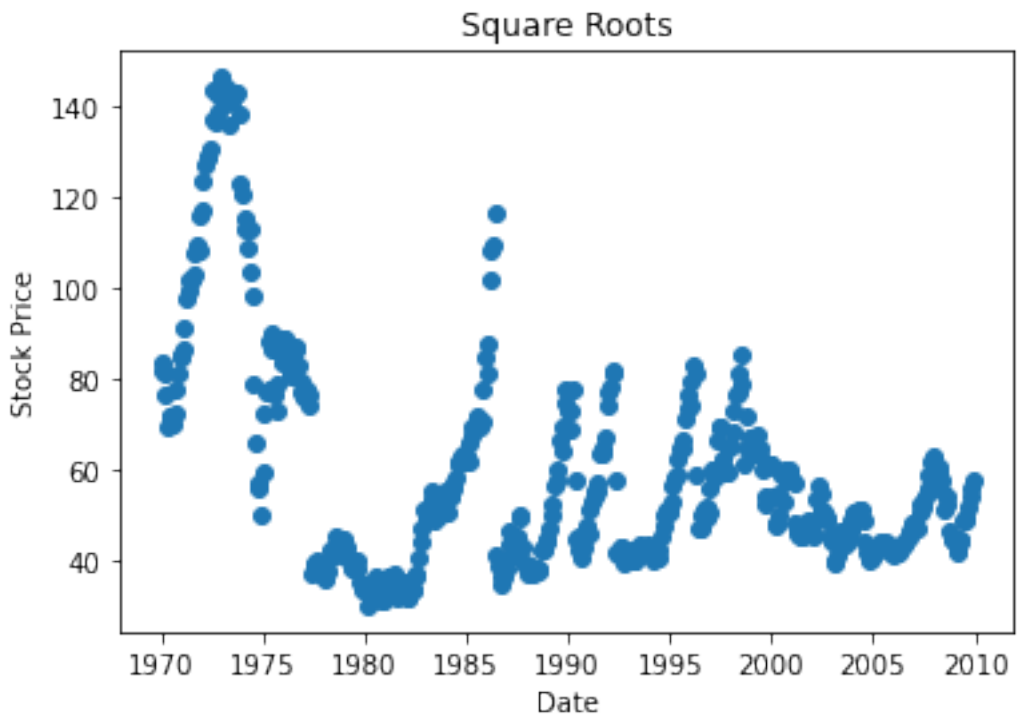
Answer - Coca-Cola Highest/Lowest stock price = **1973/1980**

What calendar year did it look to have the biggest (Year-over-Year) percentage increase?

Answer - Biggest Year over Year percentage increase = **1971 through 1973**

[104]:
```
# We can also draw a scatter draw to see which plot represent our data more
plt.xlabel('Date')
plt.ylabel("Stock Price")
plt.title("Square Roots")
plt.scatter(CocaCola.Date, CocaCola.StockPrice)
```

[104]: <matplotlib.collections.PathCollection at 0x190346847f0>

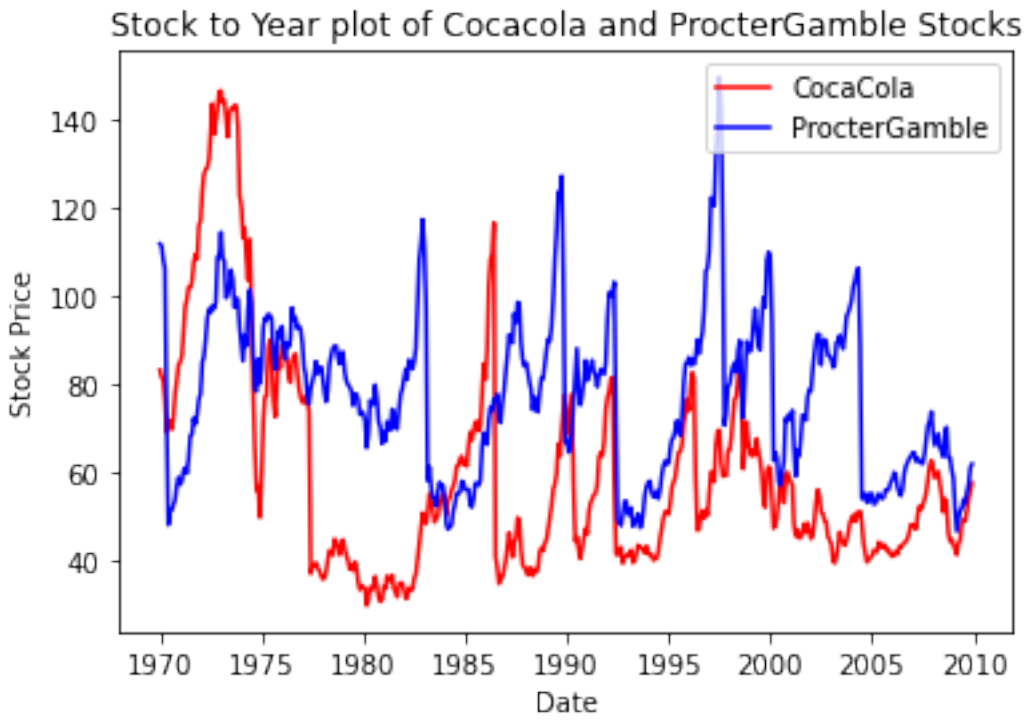


3.1.1 2. Basic Plotting Questions

Part two: Next we want to add P&G stock price onto the same graph. Go back to the plot function and add argument col="red" and col="blue" argument for CocaCola and P&G respectively.

[105]:
```
plt.xlabel('Date')
plt.ylabel("Stock Price")
plt.title("Stock to Year plot of Cocacola and ProcterGamble Stocks")
plt.plot(CocaCola.Date, CocaCola.StockPrice, 'red', label='CocaCola' )
plt.plot(ProcterGamble.Date, ProcterGamble.StockPrice, 'blue',␣
 ↪label='ProcterGamble' )
plt.legend(loc='upper right')
```

[105]:
```
<matplotlib.legend.Legend at 0x1903472fd00>
```



Based on the above line plot of Cocacola and P&G stock, we can answer the following two questions

In March of 2000 the stock market plummeted as the tech bubble burst. Using the plot above, which company's stock dropped more (relatively – i.e. percentage-wise)?

```
**Answer** -  **ProcterGamble** droped more.
```

In the year 1983 which company stock was going up? Which was going down?

```
**Answer** - **Cocacola** was going **up** and P&G was going down in 1983
```

Across the entire time period shown in your plot which stock had a generally lower price?

```
**Answer** - **Cocacola** stock had a generally low price
```

4 Data Visualization from 1995-2005

Instead of looking at the plot across the entire date range, we want to see what's happening between 1995-2005. Remember, you can use the matrix notation [rows, columns] to subset data.

First stock price of the year 1995 sits in row position: **300**

[106]:
```
IBM.loc[IBM.Date.dt.year == 1995].head(1)
```

[106]:
```
           Date  StockPrice
300 1995-01-01   74.849048
```

Last stock price of the year 2005 sits in row position: **431**

[107]:
```
IBM.loc[IBM.Date == "2005-12-01"].tail()
```

[107]:
```
           Date  StockPrice
431 2005-12-01   85.137619
```

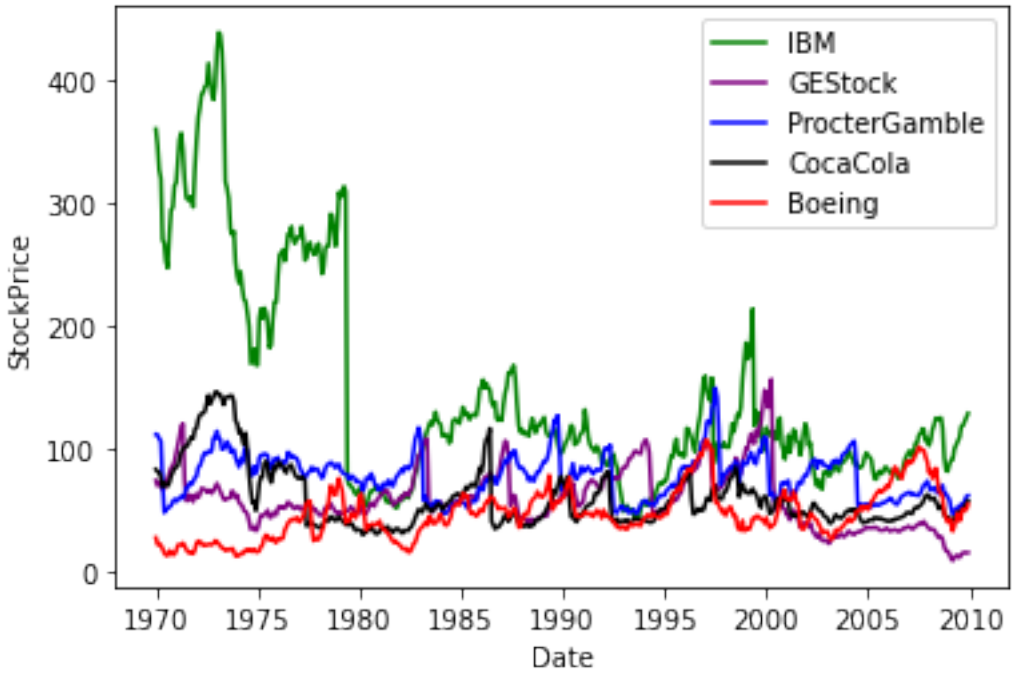
4.1 Answer the following questions:

Which stock price fell the most right after the tech bubble of March 2000? Answer: ProcterGamble fell the most after the tech bubble of March 2000

[108]:
```
# To Answer the question first we need to plot all stock on a line plot

sns.lineplot(data=IBM, x=IBM.Date, y=IBM.StockPrice, label = "IBM",␣
 ↪color="green")
sns.lineplot(data=GE, x=GE.Date, y=GE.StockPrice, label = "GEStock",␣
 ↪color="purple")
sns.lineplot(data=ProcterGamble, x=ProcterGamble.Date, y=ProcterGamble.
 ↪StockPrice, label = "ProcterGamble", color="blue")
sns.lineplot(data=CocaCola, x=CocaCola.Date, y=CocaCola.StockPrice, label =␣
 ↪"CocaCola", color="black")
sns.lineplot(data=Boeing, x=Boeing.Date, y=Boeing.StockPrice, label = "Boeing",␣
 ↪color="red")
```

[108]:
```
<AxesSubplot:xlabel='Date', ylabel='StockPrice'>
```



As anyone can see by the blue line in the plot, **Answer: ProcterGamble** has fell the most after the tech bubble of March 2000.

What stock had the highest maximum price between 1995-2005? Answer: IBM has the highest maximum price between 1995 - 2005, you can see the **Green** line in the bove diagram.

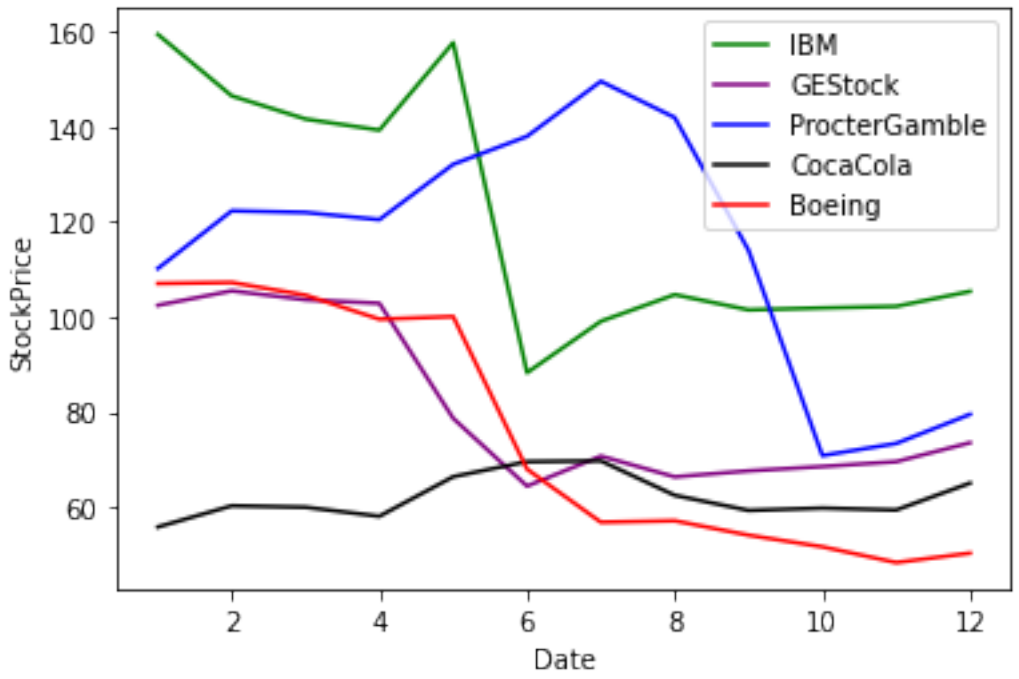
A few years before the tech bubble of 1997, there was another stock market crash trigged by economic crisis in Asia in October of 1997. If you compare stock prices from September 1997 to November 1997, which companies saw a decrease in price? Which company experienced the biggest decrease? Answer: ProcteeGamble and Boeing , have both decresesd in the specified time peried between September - and Novemeber but **Procter Gamble experienced the greatest decrease** as show by the blue line and red line consectivily on the diagram below.

[109]:
```
# To answer the question we need to sample out the exact data and draw it in a␣
↪more specific grap
aisia_crisis_IBM = IBM[IBM.Date.dt.year == 1997]
aisia_crisis_GE = GE[GE.Date.dt.year == 1997]
aisia_crisis_ProcterGamble = ProcterGamble[ProcterGamble.Date.dt.year == 1997]
aisia_crisis_CocaCola = CocaCola[CocaCola.Date.dt.year == 1997]
aisia_crisis_Boeing = Boeing[Boeing.Date.dt.year == 1997]
```

[110]:
```
```

```
sns.lineplot(data=aisia_crisis_IBM, x=aisia_crisis_IBM.Date.dt.month,␣
 ↪y=aisia_crisis_IBM.StockPrice, label = "IBM", color="green")
sns.lineplot(data=aisia_crisis_GE, x=aisia_crisis_GE.Date.dt.month,␣
 ↪y=aisia_crisis_GE.StockPrice, label = "GEStock", color="purple")
sns.lineplot(data=aisia_crisis_ProcterGamble, x=aisia_crisis_ProcterGamble.Date.
 ↪dt.month, y=aisia_crisis_ProcterGamble.StockPrice, label = "ProcterGamble",␣
 ↪color="blue")
sns.lineplot(data=aisia_crisis_CocaCola, x=aisia_crisis_CocaCola.Date.dt.month,␣
 ↪y=aisia_crisis_CocaCola.StockPrice, label = "CocaCola", color="black")
sns.lineplot(data=aisia_crisis_Boeing, x=aisia_crisis_Boeing.Date.dt.month,␣
 ↪y=aisia_crisis_Boeing.StockPrice, label = "Boeing", color="red")
```

[110]: <AxesSubplot:xlabel='Date', ylabel='StockPrice'>



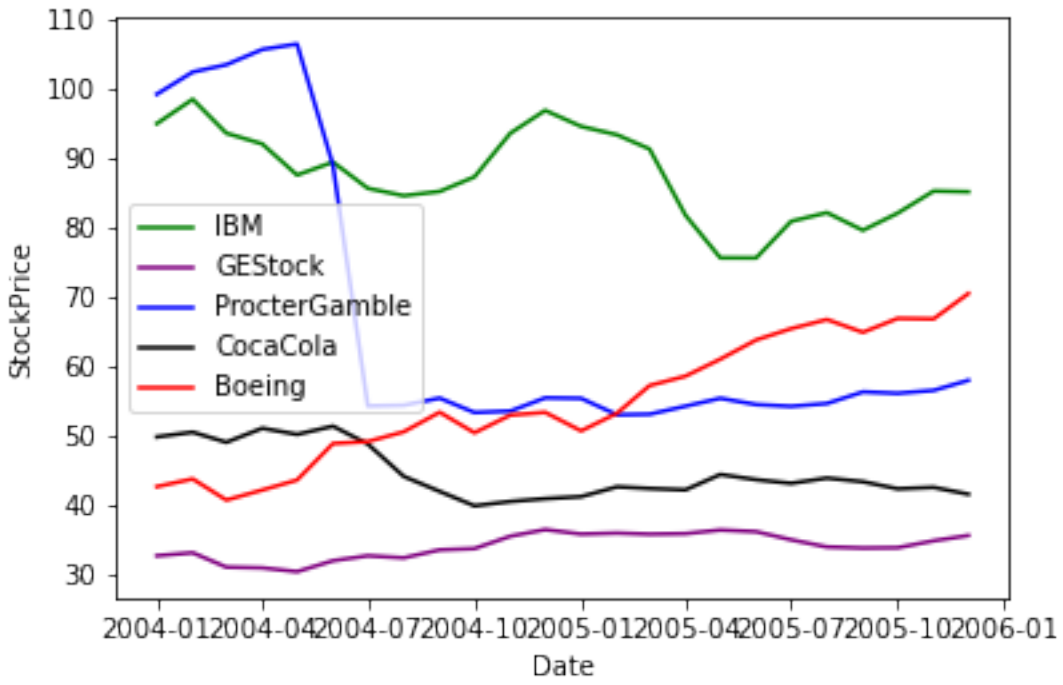
Which stock seemed to provide the best return (i.e. increase in price) between 2004-2005? Answer: Boeing , you can verify boing retunr from 2004 too the end of 2005 by the red line in the diagram below.

[111]:
```
new_ibm_data = IBM[(IBM.Date.dt.year >= 2004) & (IBM.Date.dt.year <= 2005)]
new_ge_data = GE[(GE.Date.dt.year >= 2004) & (GE.Date.dt.year <= 2005)]
new_pg =  ProcterGamble[(ProcterGamble.Date.dt.year >= 2004) & (ProcterGamble.
 ↪Date.dt.year <= 2005)]
new_coca = CocaCola[(CocaCola.Date.dt.year >= 2004) & (CocaCola.Date.dt.year <=␣
 ↪2005)]
```

```
new_boeing = Boeing[(Boeing.Date.dt.year >= 2004) & (Boeing.Date.dt.year <=␣
↪2005)]


sns.lineplot(data=new_ibm_data, x=new_ibm_data.Date, y=new_ibm_data.StockPrice,␣
↪label = "IBM", color="green")
sns.lineplot(data=new_ge_data, x=new_ge_data.Date, y=new_ge_data.StockPrice,␣
↪label = "GEStock", color="purple")
sns.lineplot(data=new_pg, x=new_pg.Date, y=new_pg.StockPrice, label =␣
↪"ProcterGamble", color="blue")
sns.lineplot(data=new_coca, x=new_coca.Date, y=new_coca.StockPrice, label =␣
↪"CocaCola", color="black")
sns.lineplot(data=new_boeing, x=new_boeing.Date, y=new_boeing.StockPrice, label␣
↪= "Boeing", color="red")
```

[111]: <AxesSubplot:xlabel='Date', ylabel='StockPrice'>



Between 1995-2005, which company had the biggest delta between the maximum and minimum stock price? Answer: IBM , has the biggest delta vale of **146.763**. Check the calculation of delta for each stock

[112]:

```
IBM_delta = IBM[(IBM.Date.dt.year >= 1995) & (IBM.Date.dt.year <=␣
↪2005)]['StockPrice'].max() - IBM[(IBM.Date.dt.year >= 1995) & (IBM.Date.dt.
↪year <= 2005)]['StockPrice'].min()
GE_delta = GE[(GE.Date.dt.year >= 1995) & (GE.Date.dt.year <=␣
↪2005)]['StockPrice'].max() - GE[(GE.Date.dt.year >= 1995) & (GE.Date.dt.year␣
↪<= 2005)]['StockPrice'].min()
ProcterGamble_delta = ProcterGamble[(ProcterGamble.Date.dt.year >= 1995) &␣
↪(ProcterGamble.Date.dt.year <= 2005)]['StockPrice'].max() -␣
↪ProcterGamble[(ProcterGamble.Date.dt.year >= 1995) & (ProcterGamble.Date.dt.
↪year <= 2005)]['StockPrice'].min()
CocaCola_delta = CocaCola[(CocaCola.Date.dt.year >= 1995) & (CocaCola.Date.dt.
↪year <= 2005)]['StockPrice'].max() - CocaCola[(CocaCola.Date.dt.year >=␣
↪1995) & (CocaCola.Date.dt.year <= 2005)]['StockPrice'].min()
Boeing_delta = Boeing[(Boeing.Date.dt.year >= 1995) & (Boeing.Date.dt.year <=␣
↪2005)]['StockPrice'].max() - Boeing[(Boeing.Date.dt.year >= 1995) & (Boeing.
↪Date.dt.year <= 2005)]['StockPrice'].min()

print("IBM Delta: {}".format(IBM_delta))
print("GE Delta: {}".format(GE_delta))
print("ProcterGamble Delta: {}".format(ProcterGamble_delta))
print("Cocacola Delta: {}".format(CocaCola_delta))
print("Boeing Delta: {}".format(Boeing_delta))
```

```
IBM Delta: 146.76306522000002
GE Delta: 133.77789473
ProcterGamble Delta: 96.62526316
Cocacola Delta: 45.67551948
Boeing Delta: 80.66904762
```

Which two companies' stock price seem to be the most correlated (i.e. move up/down together)? Answer: Cocacola and ProcterGamble , have the biggest coorlation to move together. Check the plot drawn above in question one for more detail.

5 Monthly Trend Analysis

We want to see if there are any monthly patterns (i.e. consistently higher/lower prices at various months of the year). To do for each company, we essentially want to compare "mean" by month vs the overall mean across the entire date range.

5.1 Questions:

For IBM, compare the average stock price for each month to the its overall average stock price and identify all the months for which IBM historically had a higher stock price (we call this over- indexing)? Which month over-indexed the most? Answer: January, February, March, April, May , and **February**. overindexed the most within these months.

[113]:
```
total_ibm_average = IBM.StockPrice.mean()
print("The total Average of IBM stock is {}".format(total_ibm_average))
print("The follwing months have an average greator than {}".␣
 ↪format(total_ibm_average))
ibm_month_group = IBM.groupby(IBM.Date.dt.month)
ibm_month_group.StockPrice.agg('mean').loc[ibm_month_group.StockPrice.mean() >␣
 ↪total_ibm_average]
```

```
The total Average of IBM stock is 144.3750303076664
The follwing months have an average greator than 144.3750303076664
```

[113]:
```
Date
1    150.238423
2    152.693993
3    152.432690
4    152.116824
5    151.502194
Name: StockPrice, dtype: float64
```

Repeat the function you used to solve the last question for each of the 4 remaining companies. Do any of two or more companies have their highest stock price in the same months as each other? Which companies and months does this happen for? Answer: For IBM and GE , January, February, March, April, May.

Answer: ProcterGamble and CocaCola , have increased together in **January, February.**

[114]:
```
ge_month_group = GE.groupby(GE.Date.dt.month)
ge_month_group.StockPrice.agg('mean').loc[ge_month_group.StockPrice.mean() > GE.
 ↪StockPrice.mean()]
```

[114]:
```
Date
1    62.045106
2    62.520805
3    63.150548
4    64.480092
5    60.871351
Name: StockPrice, dtype: float64
```

[115]:
```
pandg_month_group = ProcterGamble.groupby(ProcterGamble.Date.dt.month)
pandg_month_group.StockPrice.agg('mean').loc[pandg_month_group.StockPrice.
 ↪mean() > ProcterGamble.StockPrice.mean()]
```

[115]:
```
Date
1     79.617984
2     79.025755
5     77.859578
11    78.456104
12    78.296608
```

```
Name: StockPrice, dtype: float64
```

[116]:
```
coca_month_group = CocaCola.groupby(ProcterGamble.Date.dt.month)
coca_month_group.StockPrice.agg('mean').loc[coca_month_group.StockPrice.mean()␣
↪> CocaCola.StockPrice.mean()]
```

[116]:
```
Date
1    60.368487
2    60.734754
3    62.071354
4    62.688882
5    61.443581
6    60.812084
Name: StockPrice, dtype: float64
```

[117]:
```
boeing_month_group = Boeing.groupby(ProcterGamble.Date.dt.month)
boeing_month_group.StockPrice.agg('mean').loc[boeing_month_group.StockPrice.
↪mean() > Boeing.StockPrice.mean()]
```

[117]:
```
Date
2    46.892233
3    46.882076
4    47.046860
5    48.137160
6    47.385255
8    46.863107
Name: StockPrice, dtype: float64
```

What trend do you see for the months of December vs January for each company? Is there an over-arching trend that applies to all companies when comparing all historical December vs January stock prices? Answer: For All Stock Prices , increased between **December and January**.