KEY_Practice18_LineGraph_ScatterPlot

July 19, 2019

1 Practice: Line Graph and Scatter Plot

1.0.1 Line Graph

Good news! Again, we are going to use line graphs to get insight into publicly traded stocks. But this time, we want to check out that investment in which of Google or Amazon would be more profitable over long or short run!

First, import `matplotlib`, `pandas`, and `seaborn` packages

[0]:
```
# load packages
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
```

[0]:
```
# mount Google Drive
from google.colab import drive
drive.mount('/content/gdrive')
path = '/content/gdrive/My Drive/SummerExperience-master/'
```

```
Drive already mounted at /content/gdrive; to attempt to forcibly remount, call
drive.mount("/content/gdrive", force_remount=True).
```

[12]:
```
filename = path + 'SampleData/stock_price_15years.csv'
# load stock_price_15years dataset and assign it to a variable called␣
 ↪stock_price
stock_price = pd.read_csv(filename)

# print stock_price to get a better sence of data
print(stock_price)
```

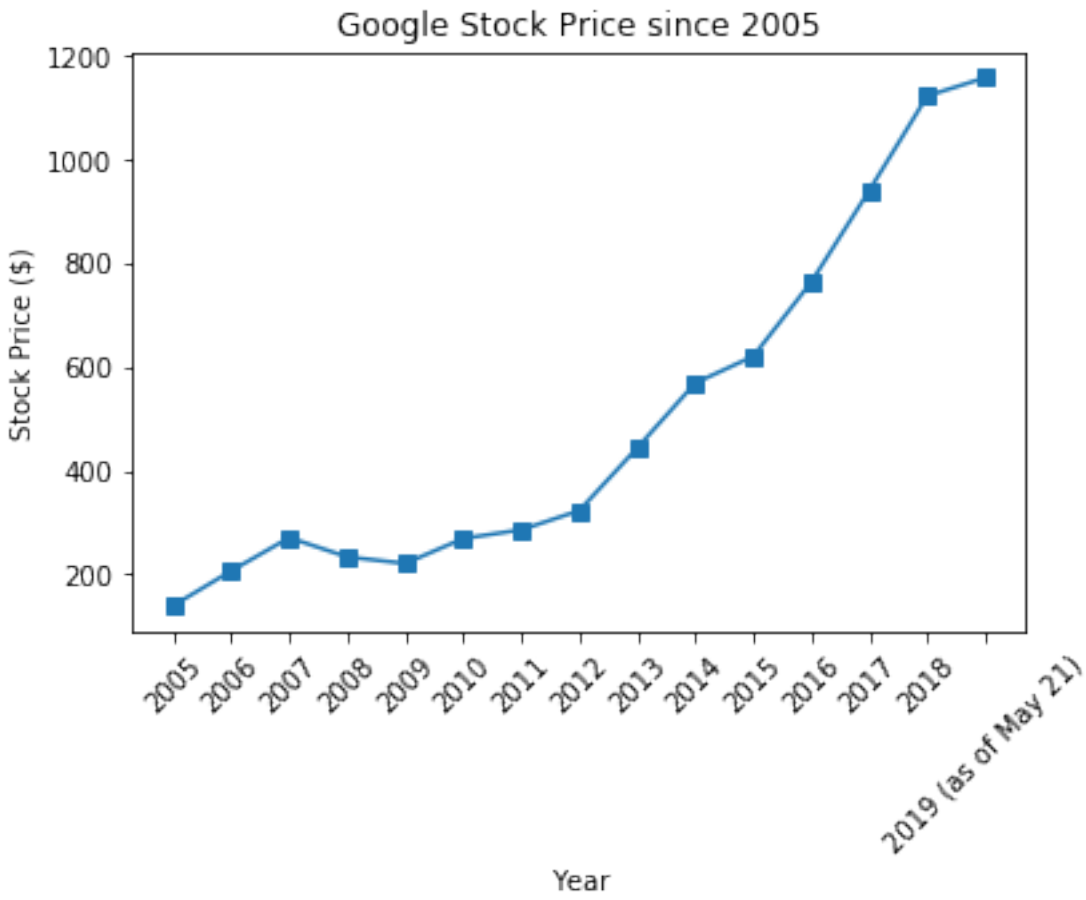
```
    year   Amazon   Google
0   2005    39.90   139.31
1   2006    35.91   206.23
2   2007    67.23   270.21
3   2008    69.88   233.15
4   2009    87.28   220.52
5   2010   139.14   268.64
```

```
6                    2011   196.67   285.37
7                    2012   220.30   322.40
8                    2013   298.03   443.49
9                    2014   332.55   568.43
10                   2015   478.14   619.98
11                   2016   699.52   763.21
12                   2017   968.17   939.77
13                   2018  1641.73  1122.04
14  2019 (as of May 21)  1742.07  1157.05
```

This dataset contains the avergae stock price of Google (Alphabet Inc.) and Amazon Inc. from 2005 to 2019 on a yearly basis.

[0]:
```
# make a line graph that shows year on the x-axis and the corresponding Google␣
 ↪stock price on the y-axis
# add labels/title/markers
# rotate the x-axis text by 45 degrees
plt.plot(stock_price["year"], stock_price["Google"], marker='s')
plt.xticks(rotation=45)
plt.title("Google Stock Price since 2005")
plt.xlabel("Year")
plt.ylabel("Stock Price ($)")
```

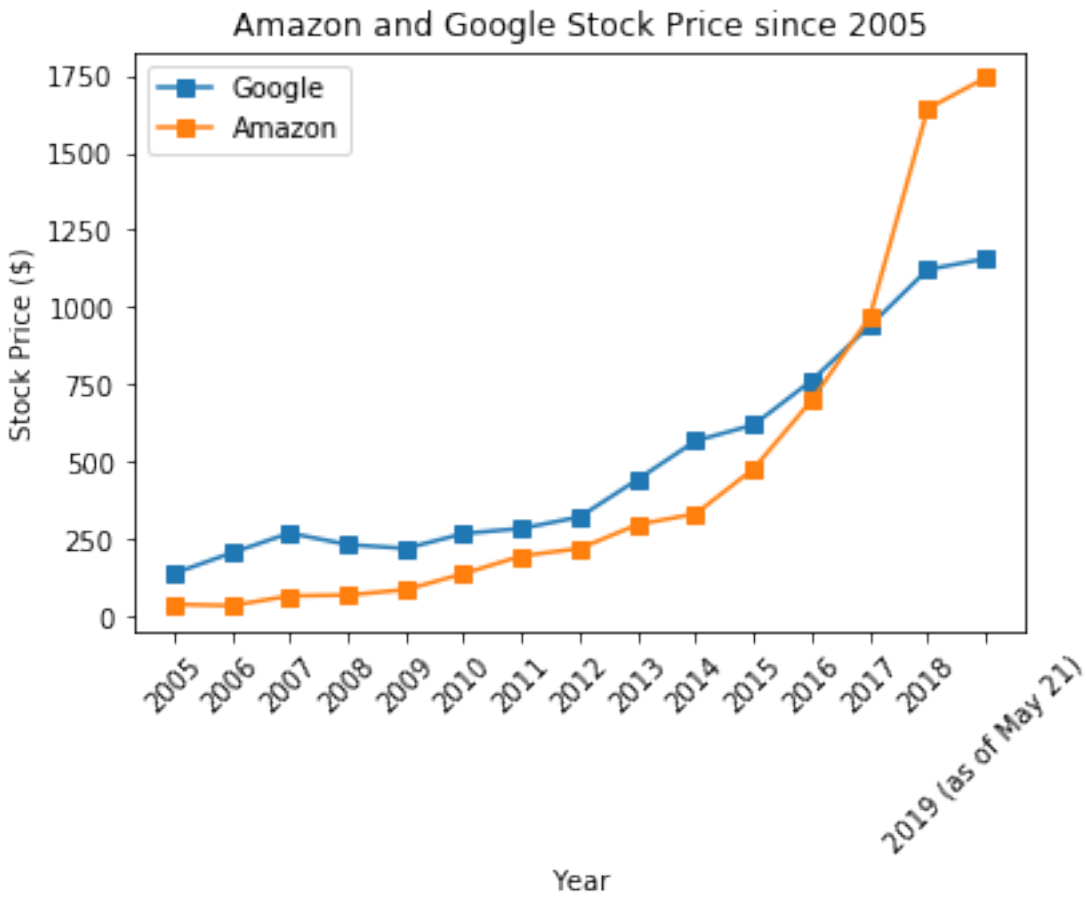
[0]:
```
Text(0, 0.5, 'Stock Price ($)')
```



Now, let's compare the average annual stock price between Amazon and Google. For that, overlay the two corresponding line graphs on the same plot.

[0]:
```
# first, plot the Google's stock price
# second, plot the Amazon's stock price
# rotate the x-axis text by 45 degrees
# don't forget to add labels/title/marker/legends to the plot
plt.plot(stock_price["year"], stock_price["Google"], marker='s', label='Google')
plt.plot(stock_price["year"], stock_price["Amazon"], marker='s', label='Amazon')
plt.legend()
plt.title("Amazon and Google Stock Price since 2005")
plt.xlabel("Year")
plt.ylabel("Stock Price ($)")
plt.xticks(rotation=45)
```

[0]:
```
([0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14],
 <a list of 15 Text xticklabel objects>)
```



Are you excited to interpret the plot?

In which year did the Amazon stock price surpass Google's stock price?

In 2005 person A invested $1000 in Google's stock while person B invested the same amount of money in Amazon's stock. If both sell their stocks after 14 years (in 2019), which one would make more profit?

How about the short-term investment? If both buy their stock in 2005 and sell them in 2006, which one would make more profit?

1.0.2 Scatter Plot

From the lesson, we learned how to generate a scatter plot which shows the relation between two variables in multiple observations. Now, we want to use this type of plot to see the relationship between flowers' petal width and sepal width.

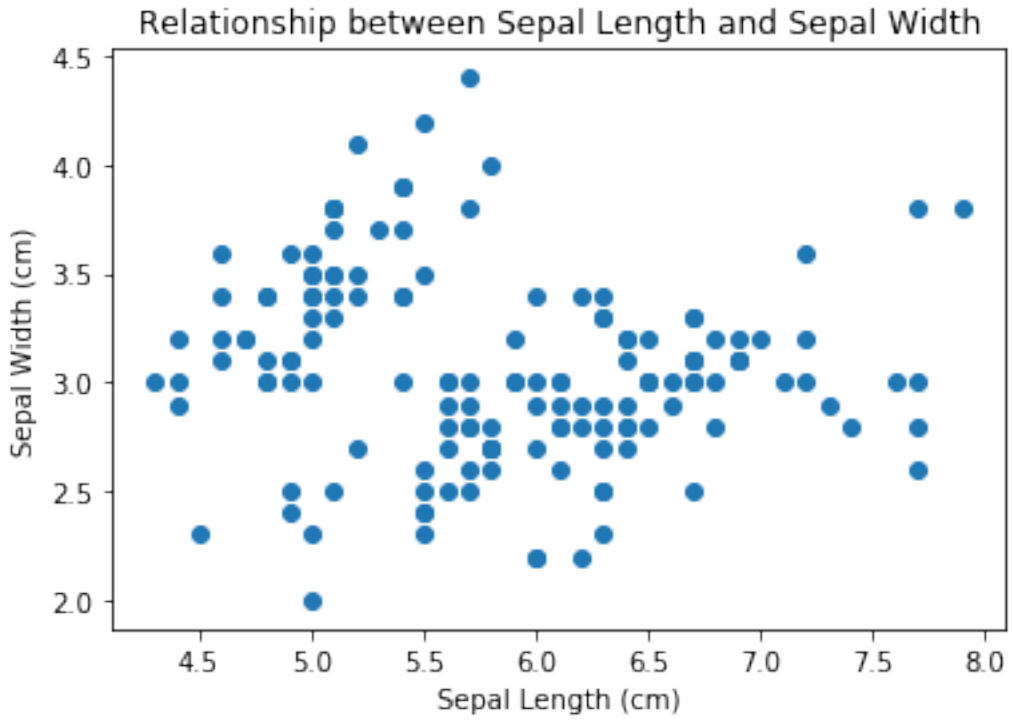
[0]:
```
# load "iris" dataset from seaborn package and assign it to a variable called␣
↪iris
# take a look at data
iris = sns.load_dataset("iris")
iris.head(10)
iris.tail(10)
```

[0]:

	sepal_length	sepal_width	petal_length	petal_width	species
140	6.7	3.1	5.6	2.4	virginica
141	6.9	3.1	5.1	2.3	virginica
142	5.8	2.7	5.1	1.9	virginica
143	6.8	3.2	5.9	2.3	virginica
144	6.7	3.3	5.7	2.5	virginica
145	6.7	3.0	5.2	2.3	virginica
146	6.3	2.5	5.0	1.9	virginica
147	6.5	3.0	5.2	2.0	virginica
148	6.2	3.4	5.4	2.3	virginica
149	5.9	3.0	5.1	1.8	virginica

[0]:
```
# generate a scatter plot that shows the petal width variable along the x-axis
# and the spetal width variable along the y-axis
# add labels/title
plt.scatter(iris["sepal_length"], iris["sepal_width"])
plt.title("Relationship between Sepal Length and Sepal Width")
plt.xlabel("Sepal Length (cm)")
plt.ylabel("Sepal Width (cm)")
```

[0]:
```
Text(0, 0.5, 'Sepal Width (cm)')
```



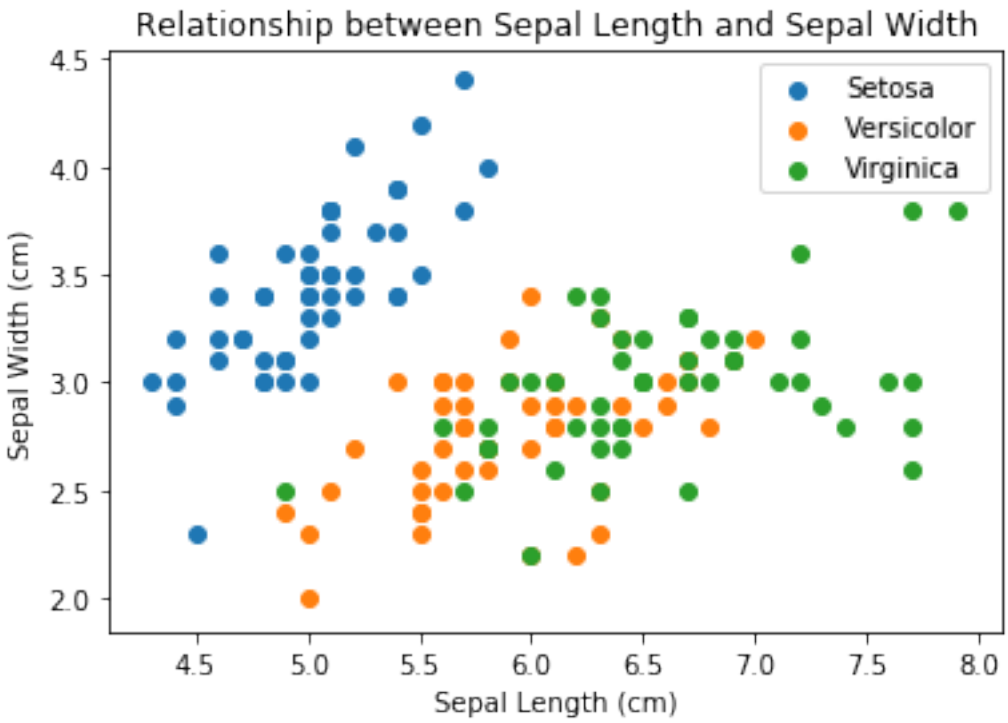
By looking at the above plot, what conclusion can you make? Is there a positive or negative correlation between sepal length and width?

One way to better understand our data would be to seperate the observations base on `species`. There are three flower species included in this study: 1) setosa , 2) versicolor, and 3) virginica. We can do this by overlaying plots, as we did in the above line graph.

[0]:
```
# generate three dataframes:
# 1) iris_setosa dataframe that only includes observations corresponding to the␣
 ↪setosa species
# 2) iris_versicolor dataframe that only includes observations corresponding to␣
 ↪the versicolor species
# 3) iris_virginica dataframe that only includes observations corresponding to␣
 ↪the virginica species
iris_setosa = iris[iris['species'] == "setosa"]
iris_versicolor = iris[iris['species'] == "versicolor"]
iris_virginica = iris[iris['species'] == "virginica"]
```

[0]:
```
# overlay three dataframes' scatter plot in one panel by:
# 1) plotting sepal width vs. length for iris_setosa
# 2) plotting sepal width vs. length for iris_versicolor
# 3) plotting sepal width vs. length for iris_virginica
# don't forget to add labels/title/legends
plt.scatter(iris_setosa["sepal_length"], iris_setosa["sepal_width"],␣
 ↪label='Setosa')
plt.scatter(iris_versicolor["sepal_length"], iris_versicolor["sepal_width"],␣
 ↪label='Versicolor')
plt.scatter(iris_virginica["sepal_length"], iris_virginica["sepal_width"],␣
 ↪label='Virginica')
plt.legend()
plt.title("Relationship between Sepal Length and Sepal Width")
plt.xlabel("Sepal Length (cm)")
plt.ylabel("Sepal Width (cm)")
```

[0]:
```
Text(0, 0.5, 'Sepal Width (cm)')
```



Now what do you notice about the relationship between these variables? Is this relationship the same for all species?

You just practiced: * creating high quality line graphs * creating high quality scatter plots * overlaying multiple dataframes in one plot, which separates them by colors

[0]: