HW2 Name: Bose Pramuanpornsatid

1

(10 points) Suppose you have a dataset $\{x\} = \{x_1, x_2, x_3, x_4\}$ consisting of 4 items. You know that $x_1 = 5$, $x_2 = 10$ and that after standardization $\hat{x}_1 = 0$, $\hat{x}_2 = 0.5$

a. Find mean $\{x\}$ and std $\{x\}$

b. Find x_3 and x_4 given that $x_3 \leq x_4$

1a)

Given $\{x\} = \{5, 10, x_3, x_4\}$

$\{\hat{x}\} = \left\{0, 0.5, \hat{x}_3, \hat{x}_4\right\}$

For any given element in standard coordinates

$$\hat{x}_i = \frac{x_i - \text{mean}(\{x_i\})}{\text{std}(\{x_i\})} \tag{1}$$

We can use formula (1) and the first element to calculate the mean.

$$\begin{aligned} \hat{x}_i &= \frac{x_i - \text{mean}(\{x_i\})}{\text{std}(\{x_i\})} \\ 0 &= \frac{5 - \text{mean}(\{x_i\})}{\text{std}(\{x_i\})} \\ 0 &= 5 - \text{mean}(\{x_i\}) \\ \text{mean}(\{x_i\}) &= 5 \end{aligned}$$

We can use formula (1), the second element, and the mean to calculate the standard deviation.

$$\begin{aligned} \hat{x}_i &= \frac{x_i - \text{mean}(\{x_i\})}{\text{std}(\{x_i\})} \\ 0.5 &= \frac{10 - 5}{\text{std}(\{x_i\})} \\ \text{std}(\{x_i\}) &= 10 \end{aligned}$$

1b)

To solve for x_3, x_4, we know mean$(\{x_i\}) = 5$ and std$(\{x_i\}) = 10$. Now we can setup a system of equation.

$$\text{mean}\,(\{x_i\}) = \frac{x_1 + x_2 + x_3 + x_4}{4} = 5$$

$$\frac{5 + 10 + x_3 + x_4}{4} = 5 \tag{2}$$

$$\frac{15 + x_3 + x_4}{4} = 5$$
$$x_3 + x_4 = 5$$

$$\text{std}(\{x_i\}) = \sqrt{\left(\frac{1}{4}\left(\sum_{i=1}^{4} (x_i - \text{mean}(\{x_i\}))^2\right)\right)} = 10$$
$$\sqrt{\left(\frac{1}{4}\left((5-5)^2 + (10-5)^2 + (x_3-5)^2 + (x_4-5)^2\right)\right)} = 10$$
$$\frac{1}{4}\left(25 + (x_3-5)^2 + (x_4-5)^2\right) = 100 \tag{3}$$
$$25 + (x_3-5)^2 + (x_4-5)^2 = 400$$
$$(x_3-5)^2 + (x_4-5)^2 = 375$$

From (2) and (3) we can solve for x_3, x_4

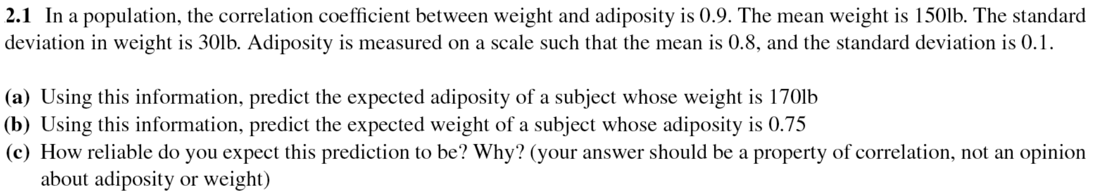
$$x_3 + x_4 = 5$$
$$(x_3-5)^2 + (x_4-5)^2 = 375$$
$$\text{------------------------}$$
$$x_4 = 5 - x_3$$
$$(x_3-5)^2 + ((5-x_3)-5)^2 = 375$$
$$(x_3-5)^2 + x_3{}^2 = 375$$
$$2\,x_3{}^2 - 10\,x_3 + 25 = 375$$
$$x_3 = \frac{5}{2} \pm \frac{5\sqrt{29}}{2} \approx -10.9629,\ 15.9629$$

Given $x_3 \le x_4$,

$x_3 \approx -10.9629$

$x_4 \approx 15.9629$

2



2.1 In a population, the correlation coefficient between weight and adiposity is 0.9. The mean weight is 150lb. The standard deviation in weight is 30lb. Adiposity is measured on a scale such that the mean is 0.8, and the standard deviation is 0.1.

(a) Using this information, predict the expected adiposity of a subject whose weight is 170lb
(b) Using this information, predict the expected weight of a subject whose adiposity is 0.75
(c) How reliable do you expect this prediction to be? Why? (your answer should be a property of correlation, not an opinion about adiposity or weight)

2a)

Let w = weight, a = adiposity Given: $\begin{array}{lll} \text{mean}(\{w\}) = 170\text{ lb}, & \text{mean}(\{a\}) = 0.8, & r = 0.9 \\ \text{std}(\{w\}) = 30\text{ lb}, & \text{std}(\{a\}) = 0.1, & \end{array}$

To predict expected adiposity for weight (w) = 170 lb. Let a_p = adiposity prediction.

$$\frac{a_p - \text{mean}(\{a\})}{\text{std}(\{a\})} = r\left(\frac{w - \text{mean}(\{w\})}{\text{std}(\{w\})}\right)$$

$$\frac{a_p - 0.8}{0.1} = 0.9\left(\frac{170 - 150}{30}\right)$$

$$a_p - 0.8 = 9\left(\frac{20}{30}\right)$$

$$a_p = 0.06 + 0.8$$

$$\text{adiposity prediction } (a_p) = \underline{0.86}$$

2b)

Let w = weight, a = adiposity Given: $\begin{array}{lll} \text{mean}(\{w\}) = 170\text{ lb}, & \text{mean}(\{a\}) = 0.8, & r = 0.9 \\ \text{std}(\{w\}) = 30\text{ lb}, & \text{std}(\{a\}) = 0.1, & \end{array}$

To predict expected adiposity for adiposity (a) = 0.75. Let w_p = weight prediction.

$$\frac{w_p - \text{mean}(\{w\})}{\text{std}(\{w\})} = r\left(\frac{a - \text{mean}(\{a\})}{\text{std}(\{a\})}\right)$$

$$\frac{w_p - 150}{30} = 0.9\left(\frac{0.75 - 0.8}{0.1}\right)$$

$$w_p - 150 = -13.5$$

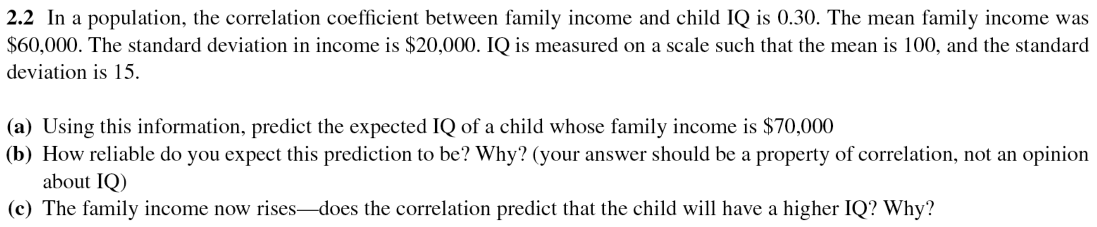
$$w_p = 150 - 13.5$$

$$\text{weight prediction } (w_p) = \underline{136.5\text{ lb}}$$

2c)

Reliability in prediction is shown by a high coefficient of determination (r^2) value of on a scale [0, 1]. In this case r^2 = 0.81 indicates that the 81% of the data should be on the regression line and only 19% of the data are inaccurate. Thus it is reliable.

3



2.2 In a population, the correlation coefficient between family income and child IQ is 0.30. The mean family income was $60,000. The standard deviation in income is $20,000. IQ is measured on a scale such that the mean is 100, and the standard deviation is 15.

(a) Using this information, predict the expected IQ of a child whose family income is $70,000
(b) How reliable do you expect this prediction to be? Why? (your answer should be a property of correlation, not an opinion about IQ)
(c) The family income now rises—does the correlation predict that the child will have a higher IQ? Why?

3a)

Let x = family income, y = IQGiven: $\begin{array}{ll} \text{mean}(\{x\}) = \$60000, & \text{mean}(\{y\}) = 100, \quad r = 0.3 \\ \text{std}(\{x\}) = \$20000, & \text{std}(\{y\}) = 15, \end{array}$

To predict expected family income (x) = $70000. Let y_p = IQ prediction.

$$\frac{y_p - \text{mean}(\{y\})}{\text{std}(\{y\})} = r\left(\frac{x - \text{mean}(\{x\})}{\text{std}(\{x\})}\right)$$

$$\frac{y_p - 100}{15} = 0.3\left(\frac{70\,000 - 60\,000}{20\,000}\right)$$

$$y_p - 100 = 4.5/2$$

$$y_p = 2.25 + 100$$

$$\text{IQ prediction}\left(a_p\right) = \underline{102.25}$$

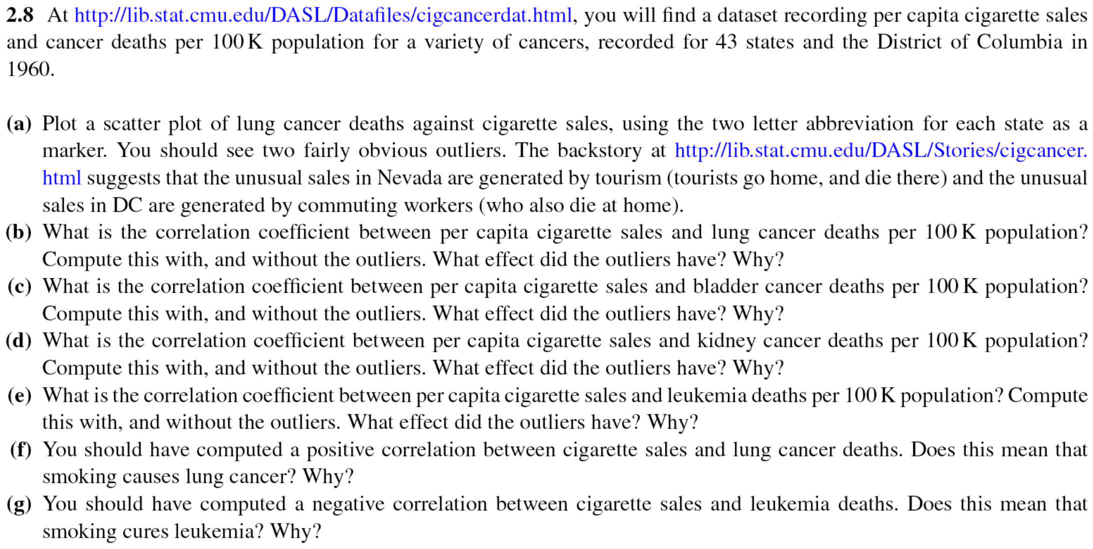
3b)

Reliability in prediction is shown by a high coefficient of determination (r^2) value of on a scale [0, 1]. In this case r^2 = 0.09 indicates that only 9% of the data should be on the regression line and only 91% of the data are inaccurate. Thus it is not reliable.

3c)

Yes, as the correlation coefficient is positive, there is a direct relationship between the two value. As one variable increase, in this case the family income, the other variable - IQ - should also rises. However correlation is not causation and we cannot be certain weather this is true, especially when the coefficient of determination is so low.

4



2.8 At http://lib.stat.cmu.edu/DASL/Datafiles/cigcancerdat.html, you will find a dataset recording per capita cigarette sales and cancer deaths per 100 K population for a variety of cancers, recorded for 43 states and the District of Columbia in 1960.

(a) Plot a scatter plot of lung cancer deaths against cigarette sales, using the two letter abbreviation for each state as a marker. You should see two fairly obvious outliers. The backstory at http://lib.stat.cmu.edu/DASL/Stories/cigcancer.html suggests that the unusual sales in Nevada are generated by tourism (tourists go home, and die there) and the unusual sales in DC are generated by commuting workers (who also die at home).

(b) What is the correlation coefficient between per capita cigarette sales and lung cancer deaths per 100 K population? Compute this with, and without the outliers. What effect did the outliers have? Why?

(c) What is the correlation coefficient between per capita cigarette sales and bladder cancer deaths per 100 K population? Compute this with, and without the outliers. What effect did the outliers have? Why?

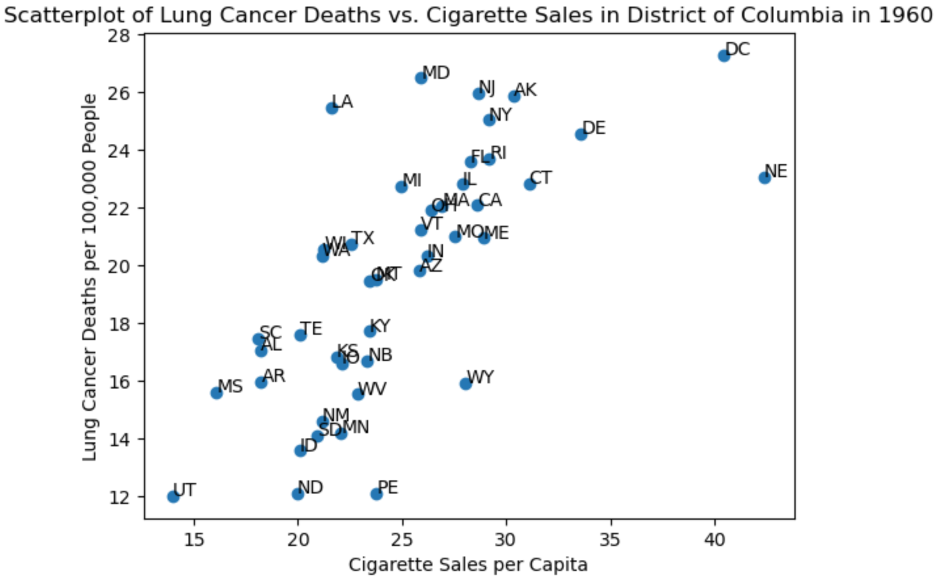
(d) What is the correlation coefficient between per capita cigarette sales and kidney cancer deaths per 100 K population? Compute this with, and without the outliers. What effect did the outliers have? Why?

(e) What is the correlation coefficient between per capita cigarette sales and leukemia deaths per 100 K population? Compute this with, and without the outliers. What effect did the outliers have? Why?

(f) You should have computed a positive correlation between cigarette sales and lung cancer deaths. Does this mean that smoking causes lung cancer? Why?

(g) You should have computed a negative correlation between cigarette sales and leukemia deaths. Does this mean that smoking cures leukemia? Why?

4a)



In[32]:=
```
import pandas as pd
import matplotlib.pyplot as plt

# Import data
data = pd.read_csv('/Users/bose/Documents/Code/UIUC/cs361-prob-stats-
cs/homework/hwk02/q4.csv')
state = data['STATE']
cig = data['CIG']
lung = data['LUNG']
```

```
# Plot data
plt.scatter(cig, lung)
plt.title('Scatterplot of Lung Cancer Deaths vs. Cigarette Sales in District of Columbia
in 1960')
plt.ylabel('Lung Cancer Deaths per 100,000 People')
plt.xlabel('Cigarette Sales per Capita')

# Annotate each point with its state abbreviation
for i in range(len(state)):
    plt.annotate(state[i], (cig[i], lung[i]))

plt.show()
```

The outliers are clearly shown in the top right: DC, NE

4b)

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

# Import data
data = pd.read_csv('q4.csv')
cig = data['CIG']
lung = data['LUNG']

# filter data with outliers
filter_wOutlier = data.filter(items=['CIG', 'LUNG'])
print(filter_wOutlier.corr())

# calculate IQR and upper/lower bounds to filter outliers
cig_iqr = cig.quantile(0.75) - cig.quantile(0.25)
lung_iqr = lung.quantile(0.75) - lung.quantile(0.25)

cig_upper = cig.quantile(0.75) + 1.5 * cig_iqr
lung_upper = lung.quantile(0.75) + 1.5 * lung_iqr

lung_lower = lung.quantile(0.25) - 1.5 * lung_iqr
cig_lower = cig.quantile(0.25) - 1.5 * cig_iqr

# filter data without outliers
filter_nOutlier = filter_wOutlier[
    (data['CIG'] < cig_upper) & (data['LUNG'] < lung_upper) &
    (data['CIG'] > cig_lower) & (data['LUNG'] > lung_lower)
]

print(filter_nOutlier.corr())
```

```
Correlation coefficient between per capita cigarette sales in District of Columbia in
1960 and lung cancer deaths per 100,000 people **With Outliers**

           CIG      LUNG
CIG   1.000000  0.697403
LUNG  0.697403  1.000000
```

```
Correlation coefficient between per capita cigarette sales in District of Columbia in
1960 and lung cancer deaths per 100,000 people **Without Outliers**

          CIG     LUNG
CIG   1.00000  0.71448
LUNG  0.71448  1.00000
```

Correlation coefficient between per capita cigarette sales in District of Columbia in 1960 and lung cancer deaths per 100,000 people **with outliers** is *r* = 0.697403.

Correlation coefficient between per capita cigarette sales in District of Columbia in 1960 and lung cancer deaths per 100,000 people **without outliers** is *r* = 0.71448.

Removing the outlier, there is a increased in the correlation coefficient because the outlier negatively impact the correlation of the data, and by removing them the prediction using correlation is more accurate.

4c)

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

# Import data
data = pd.read_csv('q4.csv')
cig = data['CIG']
blad = data['BLAD']

# filter data with outliers
filter_wOutlier = data.filter(items=['CIG', 'BLAD'])
print(filter_wOutlier.corr())

# calculate IQR and upper/lower bounds to filter outliers
cig_iqr = cig.quantile(0.75) - cig.quantile(0.25)
blad_iqr = blad.quantile(0.75) - blad.quantile(0.25)

cig_upper = cig.quantile(0.75) + 1.5 * cig_iqr
blad_upper = blad.quantile(0.75) + 1.5 * blad_iqr

blad_lower = blad.quantile(0.25) - 1.5 * blad_iqr
cig_lower = cig.quantile(0.25) - 1.5 * cig_iqr

# filter data without outliers
filter_nOutlier = filter_wOutlier[
    (data['CIG'] < cig_upper) & (data['BLAD'] < blad_upper) &
    (data['CIG'] > cig_lower) & (data['BLAD'] > blad_lower)
]

print(filter_nOutlier.corr())
```

```
Correlation coefficient between per capita cigarette sales in District of Columbia in
1960 and bladder cancer deaths per 100,000 people **With Outliers**

           CIG      BLAD
CIG   1.000000  0.703622
```

```
BLAD  0.703622  1.000000

Correlation coefficient between per capita cigarette sales in District of Columbia in
1960 and bladder cancer deaths per 100,000 people **Without Outliers**

           CIG      BLAD
CIG   1.000000  0.607626
BLAD  0.607626  1.000000
```

Correlation coefficient between per capita cigarette sales in District of Columbia in 1960 and bladder cancer deaths per 100,000 people **with outliers** is *r* = 0.703622.

Correlation coefficient between per capita cigarette sales in District of Columbia in 1960 and bladder cancer deaths per 100,000 people **without outliers** is *r* = 0.607626.

Removing the outlier, there is a decreased in the correlation coefficient because the outlier positively impact the correlation of the data, and by removing them the prediction using correlation is less accurate.

4d)

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

# Import data
data = pd.read_csv('q4.csv')
cig = data['CIG']
kid = data['KID']

# filter data with outliers
filter_wOutlier = data.filter(items=['CIG', 'KID'])
print(filter_wOutlier.corr())

# calculate IQR and upper/lower bounds to filter outliers
cig_iqr = cig.quantile(0.75) - cig.quantile(0.25)
kid_iqr = kid.quantile(0.75) - kid.quantile(0.25)

cig_upper = cig.quantile(0.75) + 1.5 * cig_iqr
kid_upper = kid.quantile(0.75) + 1.5 * kid_iqr

kid_lower = kid.quantile(0.25) - 1.5 * kid_iqr
cig_lower = cig.quantile(0.25) - 1.5 * cig_iqr

# filter data without outliers
filter_nOutlier = filter_wOutlier[
    (data['CIG'] < cig_upper) & (data['KID'] < kid_upper) &
    (data['CIG'] > cig_lower) & (data['KID'] > kid_lower)
]

print(filter_nOutlier.corr())
```

```
Correlation coefficient between per capita cigarette sales in District of Columbia in
1960 and kidney cancer deaths per 100,000 people **With Outliers**
```

```
         CIG        KID
CIG  1.00000    0.48739
KID  0.48739    1.00000

Correlation coefficient between per capita cigarette sales in District of Columbia in
1960 and kidney cancer deaths per 100,000 people **Without Outliers**

          CIG       KID
CIG  1.000000  0.548536
KID  0.548536  1.000000
```

Correlation coefficient between per capita cigarette sales in District of Columbia in 1960 and kidney deaths per 100,000 people **with outliers** is *r* = 0.48739.

Correlation coefficient between per capita cigarette sales in District of Columbia in 1960 and kidney deaths per 100,000 people **without outliers** is *r* = 0.548536.

Removing the outlier, there is a increased in the correlation coefficient because the outlier negatively impact the correlation of the data, and by removing them the prediction using correlation is more accurate.

4e)

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

# Import data
data = pd.read_csv('q4.csv')
cig = data['CIG']
luek = data['LEUK']

# filter data with outliers
filter_wOutlier = data.filter(items=['CIG', 'LEUK'])
print(filter_wOutlier.corr())

# calculate IQR and upper/lower bounds to filter outliers
cig_iqr = cig.quantile(0.75) - cig.quantile(0.25)
luek_iqr = luek.quantile(0.75) - luek.quantile(0.25)

cig_upper = cig.quantile(0.75) + 1.5 * cig_iqr
luek_upper = luek.quantile(0.75) + 1.5 * luek_iqr

luek_lower = luek.quantile(0.25) - 1.5 * luek_iqr
cig_lower = cig.quantile(0.25) - 1.5 * cig_iqr

# filter data without outliers
filter_nOutlier = filter_wOutlier[
    (data['CIG'] < cig_upper) & (data['LEUK'] < luek_upper) &
    (data['CIG'] > cig_lower) & (data['LEUK'] > luek_lower)
]

print(filter_nOutlier.corr())
```

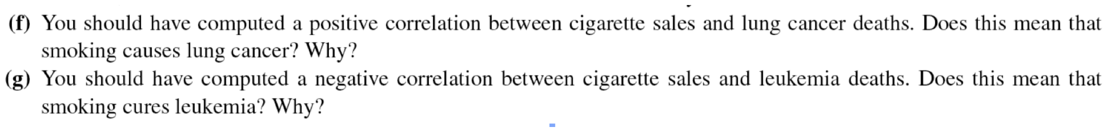
```
Correlation coefficient between per capita cigarette sales in District of Columbia in
1960 and leukemia deaths per 100,000 people **With Outliers**
```

```
           CIG      LEUK
CIG   1.000000 -0.068481
LEUK -0.068481  1.000000

Correlation coefficient between per capita cigarette sales in District of Columbia in
1960 and leukemia deaths per 100,000 people **Without Outliers**

           CIG      LEUK
CIG   1.000000  0.037071
LEUK  0.037071  1.000000
```

Correlation coefficient between per capita cigarette sales in District of Columbia in 1960 and leukemia deaths per 100,000 people **with outliers** is $r = -0.068481$.

Correlation coefficient between per capita cigarette sales in District of Columbia in 1960 and leukemia deaths per 100,000 people **without outliers** is $r = 0.037071$.

Removing the outlier, there is a increased in the correlation coefficient because the outlier negatively impact the correlation of the data, and by removing them the prediction using correlation is more accurate.



(f) You should have computed a positive correlation between cigarette sales and lung cancer deaths. Does this mean that smoking causes lung cancer? Why?

(g) You should have computed a negative correlation between cigarette sales and leukemia deaths. Does this mean that smoking cures leukemia? Why?

4f)

No, as always correlation doesn’t mean causation. This means that we cannot be certain that smoking causes lung cancer despite positive correlation.

4g)

No, as always correlation doesn’t mean causation. This means that we cannot be certain that smoking cures leukemia despite negative correlation.

5

Download the daily adjusted closing stock prices for the past 3 months of the Coca-Cola Company (KOLinks to an external site.) and PepsiCo (PEPLinks to an external site.).

a) Use this data to find the correlation coefficient between the stock prices of these two corporations

b) Plot a scatter plot with KO prices on the horizontal axis and PEP prices on the vertical axis

c) Add a prediction line to your plot that shows predictions of PEP prices from KO prices

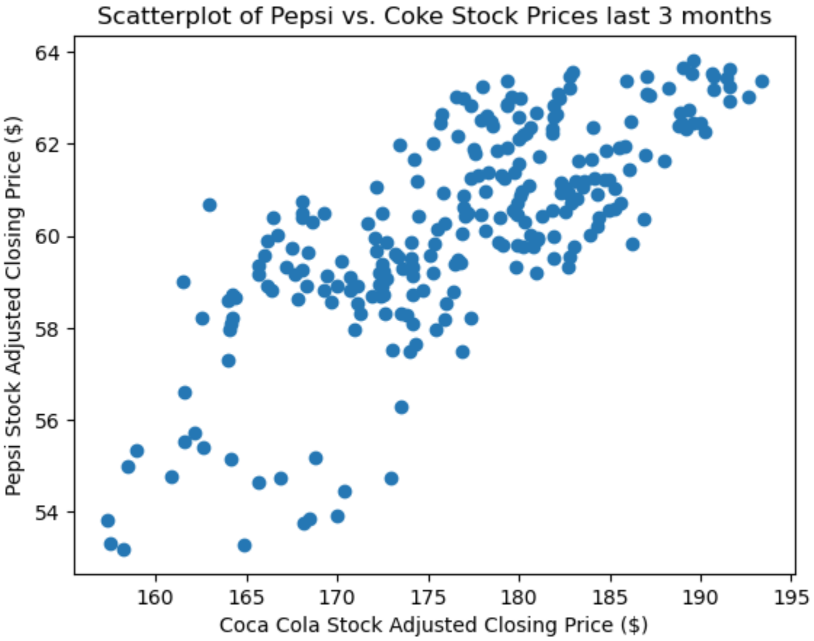
Extracted data on: 2023-09-07

5a)

```
import pandas as pd

pepsi = pd.read_csv('q5_pepsi.csv')
coke = pd.read_csv('q5_coke.csv')

print(pepsi['Adj Close'].corr(coke['Adj Close']))
```

The correlation coefficient between stock prices of Pepsi and Coca Cola is $r = 0.75437$

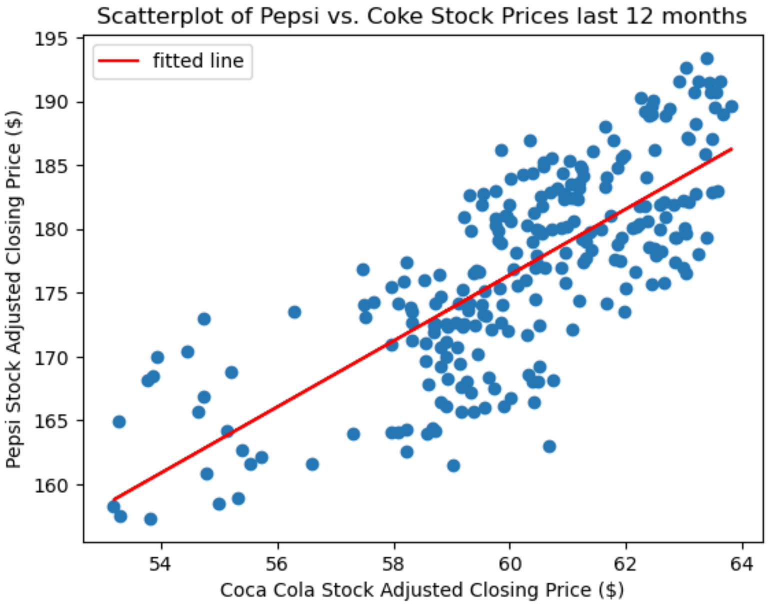
5b)



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

pepsi = pd.read_csv('q5_pepsi.csv')
coke = pd.read_csv('q5_coke.csv')
```

```
# scatter plot
plt.scatter(pepsi['Adj Close'], coke['Adj Close'])
plt.title('Scatterplot of Pepsi vs. Coke Stock Prices last 3 months')
plt.ylabel('Pepsi Stock Adjusted Closing Price ($)')
plt.xlabel('Coca Cola Stock Adjusted Closing Price ($)')
```

5c)



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

coke = pd.read_csv('q5_coke.csv')
pepsi = pd.read_csv('q5_pepsi.csv')

# scatter plot
plt.scatter(coke['Adj Close'], pepsi['Adj Close'])
plt.title('Scatterplot of Pepsi vs. Coke Stock Prices last 12 months')
plt.xlabel('Coca Cola Stock Adjusted Closing Price ($)')
plt.ylabel('Pepsi Stock Adjusted Closing Price ($)')

# plot regression line / best-fit line
from scipy import stats
slope, intercept, r_value, p_value, std_err = stats.linregress(coke['Adj Close'],
pepsi['Adj Close'])
plt.plot(coke['Adj Close'], intercept + slope*coke['Adj Close'], 'r', label='fitted
line')
plt.legend()

plt.show()
```

6

Let $\{\hat{x}_i\}$ be the standardized data set that is derived from $\{x_i\}$ and it has N items. Prove the vector $\left\langle \frac{\hat{x}_1}{\sqrt{N}}, \ldots, \frac{\hat{x}_N}{\sqrt{N}} \right\rangle$ has unit length.

$$\text{(vector magnitide)} \;\; \| v \| = \sqrt{{v_1}^2 + \ldots + {v_N}^2} \tag{4}$$

Standardization involves dividing each data point by the standard deviation, and when you sum the squares of the standardized data, you essentially get the sample size N

$$\hat{x}_1^{\,2} + \ldots + \hat{x}_N^{\,2} = N \tag{5}$$

$$\text{Given} \left\langle \frac{\hat{x}_1}{\sqrt{N}}, \ldots, \frac{\hat{x}_N}{\sqrt{N}} \right\rangle, \text{ from (4);}$$

$$\| v \| = \sqrt{\left(\frac{\hat{x}_1}{\sqrt{N}} \right)^2 + \ldots + \left(\frac{\hat{x}_N}{\sqrt{N}} \right)^2}$$

$$\text{from (5); } \| v \| = \sqrt{\frac{N}{N}} = 1$$

We have shown that $\left\langle \frac{\hat{x}_1}{\sqrt{N}}, \ldots, \frac{\hat{x}_N}{\sqrt{N}} \right\rangle$ magnitude is 1 which indicates it has unit length.