HW 2

Kaahan Motwani

Question 1

A We are given the dataset x_1, x_2, x_3, x_4 where $x_1 = 10$, $x_2 = -6$, $\hat{x}_1 = (8/7)\sqrt{2}$, and $\hat{x}_2 = 0$. In order to find the mean(μ) and standard deviation of the dataset, we will apply the formula used to standardize data, $\frac{x_i - \mu}{\sigma} = \hat{x}_i$. Applying this formula to x_2, we get the equation

$$\frac{-6 - \mu}{\sigma} = 0 \rightarrow -6 - \mu = 0$$

which we can use to find the mean, $\mu =$ **-6**.
Next, we can apply the standardization equation to x_1 with our mean, we get

$$\frac{10 - (-6)}{\sigma} = \frac{8\sqrt{2}}{7} \rightarrow 16 * 7 = (8\sqrt{2})\sigma \rightarrow \frac{112}{8\sqrt{2}} = \sigma$$

which we can use to find the standard deviation, $\sigma = \mathbf{7\sqrt{2}}$.

B Now, to find x_3 and x_4, we must apply the formulas for standard deviation and mean of a dataset. The formulas for standard deviation and mean are

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

We already know that $\mu = -6, N = 4$, and that $x_1 = 10$ and $x_2 = -6$. Plugging these values into the formula for mean, we get

$$-24 = 10 + (-6) + x_3 + x_4$$

$$-28 = x_3 + x_4$$

Next, we can use the formula for standard deviation; we already know that $\sigma = 7\sqrt{2}$, $\mu = -6, N = 4$, and $x_1 = 10$ and $x_2 = -6$. Plugging this in, we get

$$7\sqrt{2} = \sqrt{\frac{(10 - (-6))^2 + (-6 - (-6))^2 + (x_3 - (-6))^2 + (x_4 - (-6))^2}{4}}$$

Using the two equations we just created, we can now apply a system of equations to solve for the values of x_3 and x_4. Simplifying our equation for standard deviation by squaring both sides and multiplying both sides by 4, we get

$$136 = (x_3 + 6)^2 + (x_4 + 6)^2$$

Substituting our equation (from the mean formula) into the equation above, $x_3 = -x_4 - 28$, we get

$$136 = (-x_4 - 28 + 6)^2 + (x_4 + 6)^2$$

$$136 = (-x_4 - 22)^2 + x_4^2 + 12x_4 + 36$$

$$2x_4^2 + 56x_4 + 384 = 0$$

$$x_4^2 + 28x_4 + 192 = 0$$

$$(x_4 + 12)(x_4 + 16) = 0$$

$$x_4 = -12, x_4 = -16$$

Now, we have the values for x_4, but since we are given that $x_4 \geq x_3$, we know that $\mathbf{x_4 = -12}$. Thus, using and plugging this into the equation $x_3 = -x_4 - 28$, we obtain $\mathbf{x_3 = -16}$.

Question 2

A We are given the following information: the correlation coefficient (r) between weight and adiposity is 0.9. The mean weight(μ_y) is 150lb. The standard deviation in weight(σ_y) is 30lb. Adiposity is measured on a scale such that the mean is 0.8(μ_x), and the standard deviation is 0.1(σ_x). Let us use the following formula for a linear predictor to find the predicted adiposity (represented by y_i^p) of a subject whose weight is 170lb (x_i)

$$\frac{y_i^p - \mu_y}{\sigma_y} = r\frac{x_i - \mu_x}{\sigma_x}$$

Plugging in the given values into this formula, we get

$$\frac{y_i^p - 0.8}{0.1} = 0.9 * \frac{170 - 150}{30}$$

$$y_i^p = (0.1 * 0.9 * \frac{170 - 150}{30}) + 0.8$$

$$y_i^p = 0.86$$

Thus, the predicted predicted adiposity of a subject whose weight is 170lb is **0.86**.

B Next, let us use the same formula, but interchanging our x and y, for a linear predictor to find the predicted weight (represented by x_i^p) of a subject whose adiposity is 0.75 (y_i). The formula will be

$$\frac{x_i^p - \mu_x}{\sigma_x} = r\frac{y_i - \mu_y}{\sigma_y}$$

Plugging in the given values into this formula, we get

$$\frac{x_i^p - 150}{30} = 0.9 * \frac{0.75 - 0.80}{0.10}$$

$$x_i^p = (30 * 0.9 * \frac{0.75 - 0.80}{0.10}) + 150$$

$$x_i^p = 136.5$$

Thus, the predicted predicted weight of a subject whose adiposity is 0.75 is **136.5** flb.

C To determine how reliable our prediction is, let us use root-mean-square (RMS) prediction error. The formula for RMS error is $\sqrt{1 - r^2}$. Plugging our value for the correlation coefficient into this formula yields $\sqrt{1 - (0.9)^2} = \sqrt{0.19} = \mathbf{0.44}$. This is a relatively low root mean square error, so this indicates that our prediction is highly reliable, with minimal error, as indicated by the RMS error of 0.44.

Question 3

A We are given the following information: in a population, the correlation coefficient (r) between family income and child IQ is 0.30. The mean family income (μ_x) was \$60,000. The standard deviation in income (σ_x) is \$20,000. IQ is measured on a scale such that the mean is 100 (μ_y), and the standard deviation is 15 (σ_y). Let us use the following formula for a linear predictor to find the predicted IQ (represented by y_i^p) of a child whose family income is \$70,000 (x_i)

$$\frac{y_i^p - \mu_y}{\sigma_y} = r\frac{x_i - \mu_x}{\sigma_x}$$

Plugging in the given values into this formula, we get

$$\frac{y_i^p - 100}{15} = 0.30 * \frac{70000 - 60000}{20000}$$

$$y_i^p = (15 * 0.30 * \frac{70000 - 60000}{20000}) + 100$$

$$y_i^p = 102.25$$

B To determine how reliable our prediction is, let us use root-mean-square (RMS) prediction error. The formula for RMS error is $\sqrt{1 - r^2}$. Plugging our value for the correlation coefficient into this formula yields $\sqrt{1 - (0.30)^2} = \sqrt{0.91} = \mathbf{0.95}$. This is an extremely high root mean square error, so this indicates that our prediction is not very reliable, with possibility for significant error, as indicated by the RMS error of 0.95.

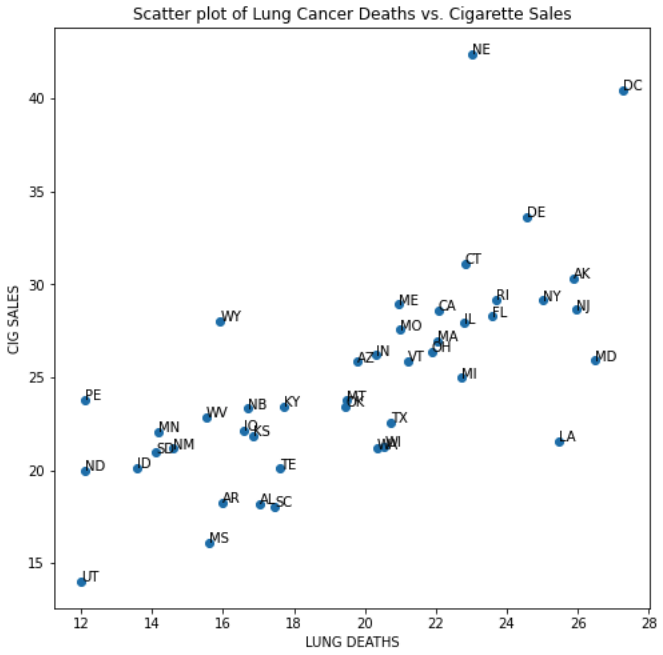
C Since the correlation is positive, the correlation does predict that the child will have a higher IQ. However, since our correlation coefficient is 0.30 and our RMS error is 0.95, the correlation is positive but not very strong and prone to error. However, since family income is positively correlated with IQ, the correlation does predict that the IQ will go up.

Question 4

A The scatter plot is shown below in **Figure 1**.

Figure 1: This is a scatter plot of lung cancer deaths against cigarette sales, with letter abbreviation markers for each state.

```
In [12]: plt.figure(figsize=(8,8))
plt.scatter(df['LUNG'], df['CIG'])
plt.xlabel("LUNG DEATHS")
plt.ylabel("CIG SALES")
plt.title("Scatter plot of Lung Cancer Deaths vs. Cigarette Sales")
for i in range(len(df)):
    plt.annotate(df.iloc[i]['STATE'], xy=(df.iloc[i]['LUNG'], df.iloc[i]['CIG']))
```



Refer to **Figure 2 (with outliers) and Figure 3 (without)** for all correlation coefficients reflected in parts B, C, D, and E.

B The correlation coefficient between per capita cigarette sales and lung cancer deaths per 100 K population, with outliers, was **0.70**. Without outliers, the correlation coefficient was **0.71**. With outliers, the correlation coefficient decreased. This is because in this case, the outliers were likely far away from both the x and y points of the rest of the dataset, so the correlation improved once they were removed.

C The correlation coefficient between per capita cigarette sales and bladder cancer deaths per 100 K population, with outliers, was **0.70**. Without outliers, the correlation coefficient was **0.61**. In this case, with outliers, the correlation coefficient increased. This is because although the outliers were very high or very low values for

one of the variables, it followed by being very high or very low for the other variable, thus improving the overall correlation coefficient.

D The correlation coefficient between per capita cigarette sales and kidney cancer deaths per 100 K population, with outliers, was **0.49**. Without outliers, the correlation coefficient was **0.58**. With outliers, the correlation coefficient decreased. This is because in this case, the outliers were likely far away from both the x and y points of the rest of the dataset, so the correlation improved once they were removed.

E The correlation coefficient between per capita cigarette sales and leukemia deaths per 100 K population, with outliers, was **-0.07**. Without outliers, the correlation coefficient was **-0.10**. In this case, with outliers, the correlation coefficient's magnitude (since correlation coefficient is negative here) decreased. This is because these two variables already have very minimal correlation, so any points that are in the top-right or bottom-left of the graph will make it seem like there is more correlation in the dataset, thus improving the correlation coefficient.

Figure 2: This is a dataframe of the correlation between all variables in this problem, **with** all outliers included.

```
df.corr()
```

	CIG	BLAD	LUNG	KID	LEUK
CIG	1.000000	0.703622	0.697403	0.487390	-0.068481
BLAD	0.703622	1.000000	0.658501	0.358814	0.162157
LUNG	0.697403	0.658501	1.000000	0.282743	-0.151584
KID	0.487390	0.358814	0.282743	1.000000	0.188713
LEUK	-0.068481	0.162157	-0.151584	0.188713	1.000000

Figure 3: This is a dataframe of the correlation between all variables in this problem, **without** outliers.

```
df_clean.corr()
```

	CIG	BLAD	LUNG	KID	LEUK
CIG	1.000000	0.607626	0.714480	0.579080	-0.101009
BLAD	0.607626	1.000000	0.640490	0.370746	0.183221
LUNG	0.714480	0.640490	1.000000	0.266764	-0.172279
KID	0.579080	0.370746	0.266764	1.000000	0.184801
LEUK	-0.101009	0.183221	-0.172279	0.184801	1.000000

F Although we computed a positive correlation between cigarette sales and lung cancer deaths, that does not mean that smoking causes lung cancer because correlation is not the same as causation. We cannot attribute any correlation between the two variables to the actual reason as to why there are lung cancer deaths occurring.

G Although we computed a negative correlation between cigarette sales and leukemia deaths, that does not mean that smoking causes lung cancer because correlation is not the same as causation. We cannot attribute any correlation (or lack thereof) between the two variables to the actual reason as to why there are leukemia deaths occurring.

Question 5

A The correlation coefficient between KO prices and PEP prices is **0.72**.

Figure 4: This is a matrix relfecting correlation coefficients between KO and PEP stock prices.

```
df_ko = pd.read_csv('KO.csv')
df_pep = pd.read_csv('PEP.csv')
del df_ko['Open'], df_ko['High'], df_ko['Low'], df_ko['Close'], df_ko['Volume']
del df_pep['Open'], df_pep['High'], df_pep['Low'], df_pep['Close'], df_pep['Volume']
df_comb = pd.DataFrame(columns=["Date", "KO Adj Close", "PEP Adj Close"])
df_comb["Date"] = df_ko["Date"]
df_comb["KO Adj Close"] = df_ko["Adj Close"]
df_comb["PEP Adj Close"] = df_pep["Adj Close"]
```

```
df_comb.corr()
```

	KO Adj Close	PEP Adj Close
KO Adj Close	1.000000	0.722468
PEP Adj Close	0.722468	1.000000

B The scatter plot is plotted below in **Figure 5**.

C The prediction line is in red in **Figure 5**.

Figure 5: This is a scatter plot of the daily adjusted close stock prices of KO and PEP.