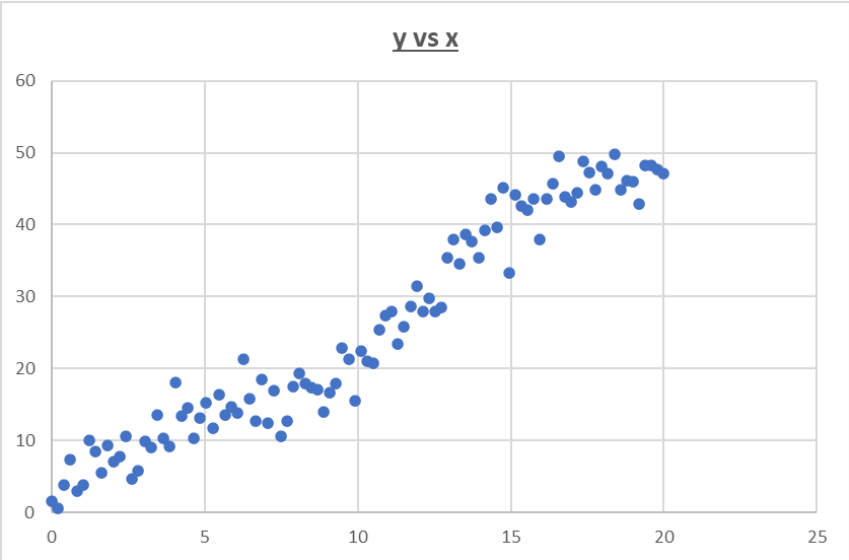
Assignment 2 - DS203

Radhika Agarwal

October 3, 2025

Section 2

Question 1



I plotted a scatter graph in an attempt to correlate the two unknown quantities, i.e. x and y. There seems to be a strong linear correlation with a small deviation from the linearity in the middle of the graph, which is in the 10-15 range where there is a dip in the data points before the upward trend is continued.

Question 2

The **Pearson coefficient** is used to judge how strong the correlation between two variables are and the direction of their relation. The formula for calculating the Pearson coefficient is the following:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}}$$

It can be intuitively written as

$$r = \frac{\mathrm{Cov}(x,y)}{\sigma_x \sigma_y}$$

The formuala makes sense because the covariance is just $\mathrm{Cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ which means that if both x and y are above their mean at the same time, the product is positive and vice versa. And the product of the standard deviations is just used to normalize the value and put it between -1 and 1. For the data above the pearson's coefficient is equal to **0.9674** The formula I used was:

```
=(100*SUMPRODUCT(A2:A101,B2:B101)-SUM(A2:A101)*SUM(B2:B101))/
SQRT((100*SUMPRODUCT(A2:A101,A2:A101)-SUM(A2:A101)^2)*
     (100*SUMPRODUCT(B2:B101,B2:B101)-SUM(B2:B101)^2))
```

Its 0.96 which is pretty close to 1 which indicates a strong positive correlation as predicted in the earlier question.

Question 3

In order to create the data sets with standard deviation higher than the current standard deviation the formula given was:

$$y_{\text{new}} = y + \text{noise}, \quad \text{where noise} \sim N(0, \sigma_{\text{noise}})$$

To write the noise in excel I have used the formula

```
NORM.INV(RAND(), 0, (value of the noise))
```

where we are creating a random number and using that number to find a corresponding value from a normal distribution with a mean of 0 and a standard deviation of the noise calculated.
So to proceed further, i have made 5 different sheets in the excel workbook where i am working the data out for each set differently. The training to testing split is 80:20 and it will be unfair to choose the first 80 rows hence to choose the rows and split data, the method opted for is that i will be creating a random number next to the data for each row and then sort the rows based on that random number generated. RAND() is a voltaile function thus i had to copy paste values at the end to avoid changes.
Now in order to get the parameter values, i have used the inbuilt function of excel known as the **LINEST** function that immediately evaluates to 10 different parameters related to the dataset. Below is the summarised results of all the parameters asked for.

Desired Std Dev	Data Split	R^2	F-statistic	RMSE
5	Train	0.938	1191.99	1.25
	Test	0.928	-	1.35
10	Train	0.928	1015.46	2.67
	Test	0.976	-	2.03
Original	Train	0.942	1275.77	3.73
	Test	0.911	-	4.26
20	Train	0.495	36.42	14.67
	Test	0.38	-	12.73
25	Train	0.318	95.3	22.14
	Test	0.445	-	20.98

Analysis

The scatter plots and the table provides a powerful demonstration of how noise (how varied the data is) affects the linear model. As the standard deviation increases the clear linear relationship becomes not so clear and eventually is just a bunch of points scattered around with no defining correspondence whatsoever. This can be complemented by the RMSE values skyrocketing while the R^2 and F values falling.
There is a clear fall of the R^2 values. For the low-noise datasets, the R^2 values are all excellent, consistently above 0.91. This means the model can explain over 91% of the variability in 'y', indicating a very strong fit. Howvever later it falls to a 0.318 which means that the model can only explain less than a third of the variance. Thus we can also conclude from this that R^2 is highly sensitive to noise explaining how it gets tougher to figure out the pattern and make accruate predictions as the noise in the data increases.

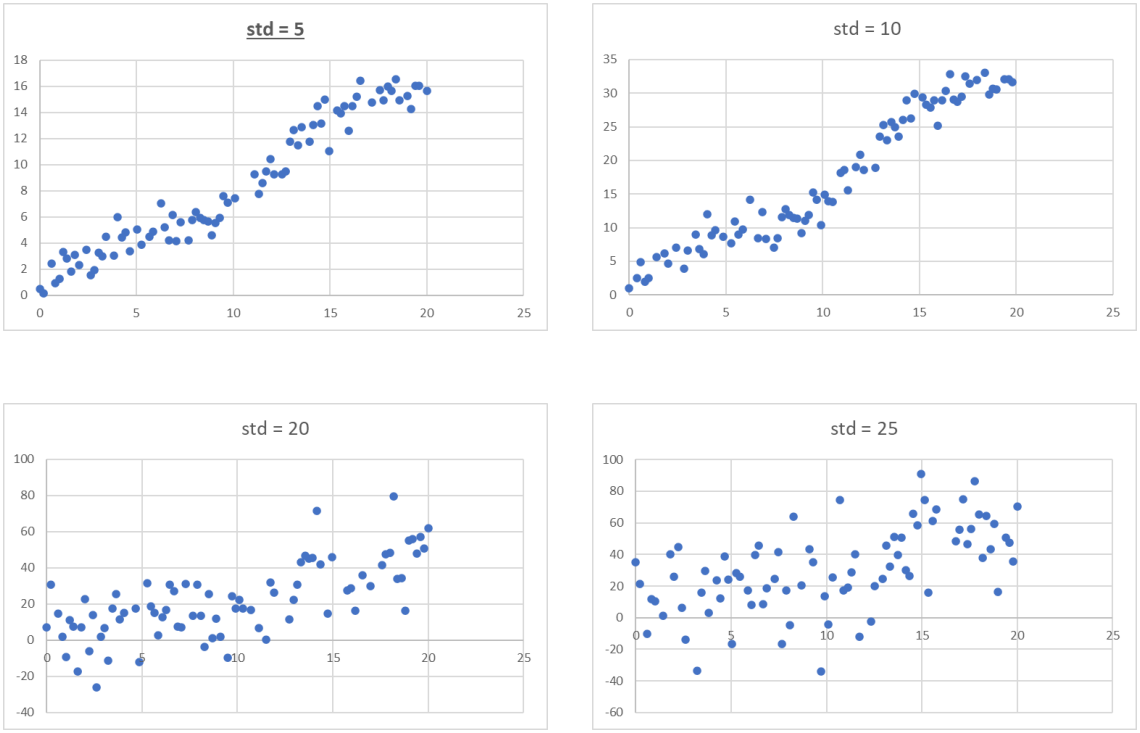


Figure 1: Shows all plots of the 4 new data sets created

Similarly for RMSE there is a clear and significant upward trend the standard deviation increases. Since RMSE measures the error, it makes sense that the value increases as our attempt to find the best fit line becomes vaguer and less precise.
Lastly the F stastic indicates how confidence we are about out model and that it can detetct clean data. And with that logic it makes sense why the F statistic value decreased drastically from a 1000 to 36.
The normal expected behaviour for the test data vs train data is that the test R^2 would be slightly lower than the train R^2 and the test RSME would be higher. And we see the first two data sets following that. The closeness of the values shows that our model is working nicely else there would have been a major difference in the set of values. Howvever the last three data sets we see the opposite happening that is the test data performed better than the training data. This is a mere fluke because the data points were chosen randomly and it just so happened that the points chosen for the test data lay closer to the best fit line.