SOEN 6611 Software Measurement

ID: 40193024
Mrinal Rai

Assignment - 2

Submitted To: **Dr. Olga Ormandjieva**

Part 1:

The below data is attached in the submission (***A2-data-2022-solution.xslx***)

Student #	Programming Langugage	SLOC: Manual counting	Effort (in minutes) to write the program
P1	Java	20	34
P2	Java	18	25
P3	java	26	30
P4	Java	53	180
P5	Java	17	11
P6	Java	19	11
P7	Java	15	19
P8	Java	20	30
P9	Java	39	20
P10	Java	15	15
P11	Java	30	35
P14	Java	15	20
P15	java	20	45
P17	java	42	60
P18	java	26	10
P19	javascript	26	45
P20	java	49	90
P21	JavaScript	115	215
P22	java	42	120
P24	java	10	20
P25	java	31	30
P29	java	37	45
P30	java	51	63
P38	java	51	60
P39	java	18	20
P40	java	28	60
P41	java	73	15
P42	java	26	30

P43	java	23	60
P44	java	280	45
P45	java	23	40
P46	java	26	39
P47	java	26	28
P48	java	26	63
P50	java	56	50
P51	java	280	45
P54	java	30	30
P55	javascript	14	70
P58	java	56	28
P59	java	31	36
P60	Java	20	30
P62	java	21	19.5
P65	java	17	30
P66	java	36	37
P68	java	50	90
P71	java	38	45
P73	java	14	25
P75	java	31	40
P77	java	26	36
P78	Java	32	40
P79	java	54	65
P83	java	25	53
P85	java	50	60
P86	java	53	70

1.1a) Averaging – mean, median, standard deviation

Mean:

List of Length (Sloc):
20, 18, 26, 53, 17, 19, 15, 20, 39, 15, 30, 15, 20, 42, 26, 26, 49, 115, 42, 10, 31, 37, 51, 51, 18, 28, 73, 26, 23, 280, 23, 26, 26, 26, 56, 280, 30, 14, 56, 31, 20, 21, 17, 36, 50, 38, 14, 31, 26, 32, 54, 25, 50, 53

Mean = Total sum of all length values / Total Count
= 2270 / 54
Mean = 42.037.

Median:

$$\mathrm{Med}(X) = \begin{cases} \frac{X[\frac{n}{2}]+X[\frac{n+1}{2}]}{2} & \text{if n is even} \\ X[\frac{n+1}{2}] & \text{if n is odd} \end{cases}$$

X = ordered list of values in data set
n = number of values in data set

X:
10, 14, 14, 15, 15, 15, 17, 17, 18, 18, 19, 20, 20, 20, 20, 21, 23, 23, 25, 26, 26, 26, 26, 26, 26, 26, 26, 28, 30, 30, 31, 31, 31, 32, 36, 37, 38, 39, 42, 42, 49, 50, 50, 51, 51, 53, 53, 54, 56, 56, 73, 115, 280, 280

n = 54

Median = (X[54/2] + X[54/2 + 1]) / 2
= (X[27] + X[28]) / 2
= 26+28 / 2
Median = 27

Standard Deviation:

σ: **49.9766475**

Count, N: 54
Sum, Σx: 2270
Mean, μ: 42.037
Variance, σ^2: 2497.665295

Steps:

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

$$\sigma 2 = \frac{\Sigma(xi - \mu)2}{N}$$

$$= \frac{(20 - 42.037)\text{\textasciicircum}2 + \ldots + (53 - 42.037)\text{\textasciicircum}2}{54}$$

$$= \frac{134873.9259}{54}$$

= 2497.665295

σ = √2497.665295
= 49.9766475

Standard Deviation = 49.9766475

Interpretation:

The mean obtained via above calculation is 42.037. It means that on the basis of historical data given to us, it takes 42 lines of code to write a program to calculate distance between two geographical coordinates on earth. Mean is highly impacted by outliers and we can see in our data set that some values like 115, 280 and 280 are high values which skewed our results.

Median helps us partition our data set in two parts. Our median is 27, which divides our datasets into two partitions.

Standard deviation helps us understand the distance of our datapoints from the mean. In our case the standard deviation is of 49.9766475 which is quite high. Hence, we cannot expect the data points to be close to our mean value which is 42.037 SLOC. This is due to availability of outliers in our dataset.

1.1 b) Box plot

Step 1 Collect Data from excel sheet for 54 programmers.

Step 2 Find median of the data set which is 27 SLOC.

Step 3 Lower and Upper Quartiles

Sorted List:
10, 14, 14, 15, 15, 15, 17, 17, 18, 18, 19, 20, 20, 20, 20, 21, 23, 23, 25, 26, 26, 26, 26, 26, 26, 26, 26, 28, 30, 30, 31, 31, 31, 32, 36, 37, 38, 39, 42, 42, 49, 50, 50, 51, 51, 53, 53, 54, 56, 56, 73, 115, 280, 280

First, we need to calculate Lower and upper forths.

Lower Forth = 1 *(n +1) /4 where n= total number of count for particular measure (Length in this case).

Lower Forth = (54+1)/4 = 14th position
Lower Forth = 20

Upper Forth = 3*(n +1) /4 where n= total number of count for particular measure (Length in this case).

Upper Forth = (3 * 55) / 4 = 41st position
Upper Forth = 49

Step 4 Box length calculation: difference between lower and upper quartile.

Box Length = Upper Forth – Lower Forth
Box Length = 49-20 = 29

Step 5 Calculation of Upper and Lower Tails

1. Multiplying the box length by 1.5
2. Adding and subtracting the box length from upper and lower tail respectively.

29*1.5 = 43.5

Upper tail = Upper forth + 43.5

Lower tail = Lower forth – 43.5

Upper tail = 49 + 43.5
Upper Tail = 92.5.

Lower Tail = 20-43.5
Lower Tail = 0

Range of Acceptable values:
It is equal to the range of [Lower Forth, Upper Forth]

[20 to 49] Values: (20, 20, 21, 23, 23, 25, 26, 26, 26, 26, 26, 26, 26, 26, 28, 30, 30, 31, 31, 31, 32, 36, 37, 38, 39, 42, 42, 49)

Range of values that need a quick review:
[Lower tail .. lower forth [U] upper forth .. upper tail]

[0...20] U [49...92.5] Values: (10, 14, 14, 15, 15, 15, 17, 17, 18, 18, 19, 20, 20, 50, 50, 51, 51, 53, 53, 54, 56, 56, 73)

Range for Outliers:
Greater than the Upper tail and lower than the lower tail
X > 92.5 Values: (115, 280, 280)
X < 0 Values: None

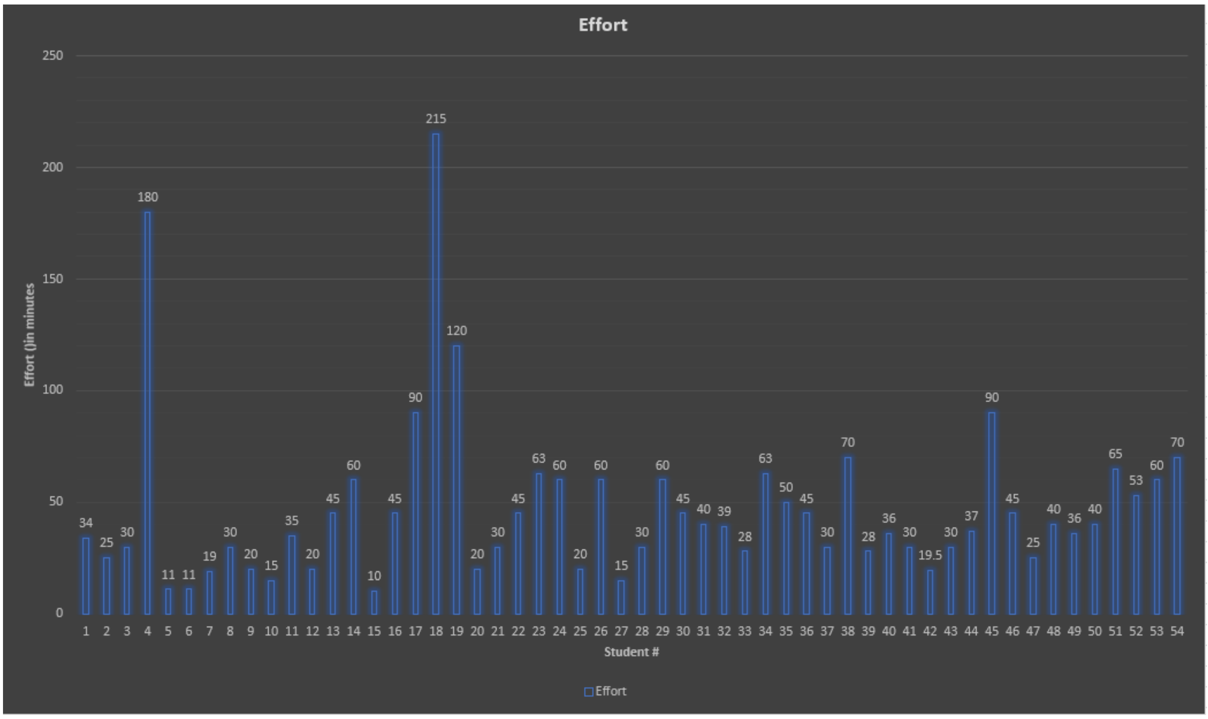
Outliers: (115, 280, 280)

Interpretation:

The box plot analysis concludes that for SLOC variable, the accepted range of values lie between lower quartile and upper quartile which is 20 to 49 respectively. Whereas the outliers that is the values residing outside the lower tail and upper tail areas (115, 280 and 280) requires stringent review and analysis as they are not acceptable. The values that need a quick review are between 0 to 20 and 49 to 92.5 as they require quick review before being accepted.

1.2) Apply Bar chart analysis technique

The below chart is attached in the submission (***A2-data-2022-solution.xslx***)



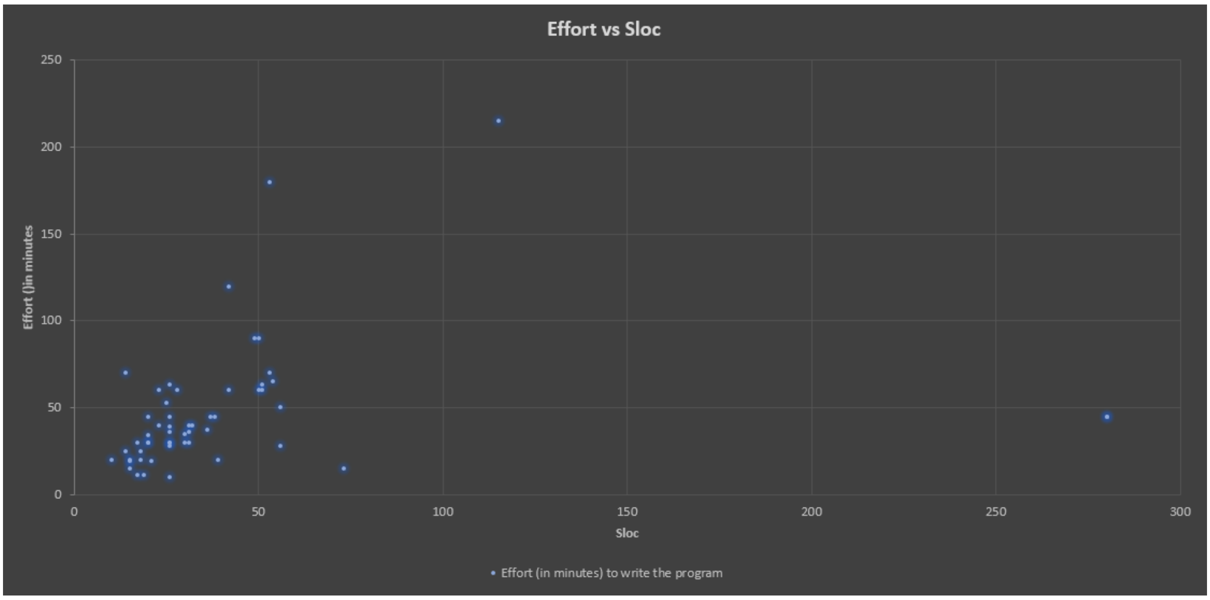
Interpretation:

It can be observed via the bar chart that on average the effort in minutes lies between 45 to 50 minutes. There are some outliers as it took some programmers 180, 215 and 120 minutes to complete the program. These values require stringent review and further analysis.

Part 2: Investigate the relationship across two variables in A2-data-2022

2.1 a) Scatter Plot

The below chart is attached in the submission (***A2-data-2022-solution.xslx***)



It can be observed in the above scatter plot that there is a relation between effort and length. In most of the cases as the length increases, the effort increases as well. So they are typically directly proportional to each other.
But it can be seen in some cases such as #P41 (73 SLOC and 15 Effort in mins) and #P55 (14 SLOC and 70 effort in mins) where this relationship does not exist and are not organized in the same way as other data points.

b) Correlation analysis

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

xi = SLOC
x̄ = Mean SLOC = 42.03703704
yi = Effort
ȳ = Mean Effort = 46.89814815

(xi - x̄) = -22.037, -24.037, -16.037, 10.963, -25.037, -23.037, -27.037, -22.037, -3.037, -27.037, -12.037, -27.037, -22.037, -0.037, -16.037, -16.037, 6.963, 72.963, -0.037, -32.037, -11.037, -5.037, 8.963, 8.963, -24.037, -14.037, 30.963, -16.037, -19.037, 237.963, -19.037, -16.037, -16.037, -16.037, 13.963, 237.963, -12.037, -28.037, 13.963, -11.037, -22.037, -21.037, -25.037, -6.037, 7.963, -4.037, -28.037, -11.037, -16.037, -10.037, 11.963, -17.037, 7.963, 10.963
(yi - ȳ) = -12.89814815, -21.89814815, -16.89814815, 133.1018519, -35.89814815, -35.89814815, -27.89814815, -16.89814815, -26.89814815, -31.89814815, -11.89814815, -26.89814815, -1.89814815, 13.10185185, -36.89814815, -1.89814815, 43.10185185, 168.1018519, 73.10185185, -26.89814815, -16.89814815, -1.89814815, 16.10185185, 13.10185185, -26.89814815, 13.10185185, -31.89814815, -16.89814815, 13.10185185, -1.89814815, -6.89814815, -7.89814815, -18.89814815, 16.10185185, 3.10185185, -1.89814815, -16.89814815, 23.10185185, -18.89814815, -10.89814815, -16.89814815, -27.39814815, -16.89814815, -9.89814815, 43.10185185, -1.89814815, -21.89814815, -6.89814815, -10.89814815, -6.89814815, 18.10185185, 6.10185185, 13.10185185, 23.10185185

Σ (xi - x̄) (yi - ȳ) = 284.2364908+ 526.3657871+ 270.9956019+ 1459.195602+ 898.7819352+ 826.9856389+ 754.2822315+ 372.3844908+ 81.68967593+ 862.4302315+ 143.2180093+ 727.2452315+ 41.82949078+ -0.484768518+ 591.7356019+ 30.44060188+ 300.1181944+ 12265.21542+ -2.704768518+ 861.7359723+ 186.5048611+ 9.560972232+ 144.3208981+ 117.4318981+ 646.5507871+ -183.9106944+ -987.6623612+ 270.9956019+ -249.4199537+ -451.6890282+ 131.3200463+ 126.6626019+ 303.0696019+ -258.2253981+ 43.31115738+ -451.6890282+ 203.4030093+ -647.7066203+ -263.8748426+ 120.2828611+ 372.3844908+ 576.3748426+ 423.0789352+ 59.75512038+ 343.2200463+ 7.662824082+ 613.9583797+ 76.13486113+ 174.7736019+ 69.23671298+ 216.5524537+ -103.95725+ 104.3300463+ 253.2656018

Σ (xi - x̄) (yi - ȳ) = 23291.7037

Σ (xi - x̄)² = 485.629369+ 577.777369+ 257.185369+ 120.187369+ 626.851369+ 530.703369+ 730.999369+ 485.629369+ 9.223369+ 730.999369+ 144.889369+ 730.999369+ 485.629369+ 0.001369+ 257.185369+ 257.185369+ 48.483369+ 5323.599369+ 0.001369+ 1026.369369+ 121.815369+ 25.371369+ 80.335369+ 80.335369+ 577.777369+ 197.037369+ 958.707369+ 257.185369+ 362.407369+ 56626.38937+ 362.407369+ 257.185369+ 257.185369+ 257.185369+ 194.965369+ 56626.38937+ 144.889369+ 786.073369+ 194.965369+ 121.815369+ 485.629369+ 442.555369+ 626.851369+ 36.445369+ 63.409369+ 16.297369+ 786.073369+ 121.815369+ 257.185369+ 100.741369+ 143.113369+ 290.259369+ 63.409369+ 120.187369

Σ (xi - x̄)² = 134873.9259

Σ (yi - ȳ)^2 = 166.3622257+ 479.5288924+ 285.5474109+ 17716.10297+ 1288.677041+ 1288.677041+ 778.3066702+ 285.5474109+ 723.5103739+ 1017.491855+ 141.5659294+ 723.5103739+ 3.602966399+ 171.6585219+ 1361.473337+ 3.602966399+ 1857.769633+ 28258.2326+ 5343.880744+ 723.5103739+ 285.5474109+ 3.602966399+ 259.269633+ 171.6585219+ 723.5103739+ 171.6585219+ 1017.491855+ 285.5474109+ 171.6585219+ 3.602966399+ 47.5844479+ 62.3807442+ 357.1400035+ 259.269633+ 9.621484899+ 3.602966399+ 285.5474109+ 533.6955589+ 357.1400035+ 118.7696331+ 285.5474109+ 750.658522+ 285.5474109+ 97.9733368+ 1857.769633+ 3.602966399+ 479.5288924+ 47.5844479+ 118.7696331+ 47.5844479+ 327.6770404+ 37.232596+ 171.6585219+ 533.6955589

Σ (yi - ȳ)^2 = 72791.68981

Sqrt(Σ (xi - x̄)^2 * Σ (yi - ȳ)^2) = 99084.31248

r = Σ (xi - x̄) (yi - ȳ) / Sqrt(Σ (xi - x̄)^2 * Σ (yi - ȳ)^2)
= 23291.7037 / 99084.31248
= 0.2351

As value is just greater than 0 which means that the correlation between length and effort is moderate.

c) Regression analysis

$Y = \beta_0 + \beta_1 X$

$\beta_1 = (\Sigma x_i * y_i) - (n*x_{avg}*y_{avg}) / (\Sigma x_i^2) - (nx_{avg}^2)$
$x_{avg} = 42.03703704$
$y_{avg} = 46.89814815$
n= 54
$\Sigma\, x_i * y_i$ = 680+ 450+ 780+ 9540+ 187+ 209+ 285+ 600+ 780+ 225+ 1050+ 300+ 900+ 2520+ 260+ 1170+ 4410+ 24725+ 5040+ 200+ 930+ 1665+ 3213+ 3060+ 360+ 1680+ 1095+ 780+ 1380+ 12600+ 920+ 1014+ 728+ 1638+ 2800+ 12600+ 900+ 980+ 1568+ 1116+ 600+ 409.5+ 510+ 1332+ 4500+ 1710+ 350+ 1240+ 936+ 1280+ 3510+ 1325+ 3000+ 3710
$\Sigma\, x_i * y_i = 129750.5$

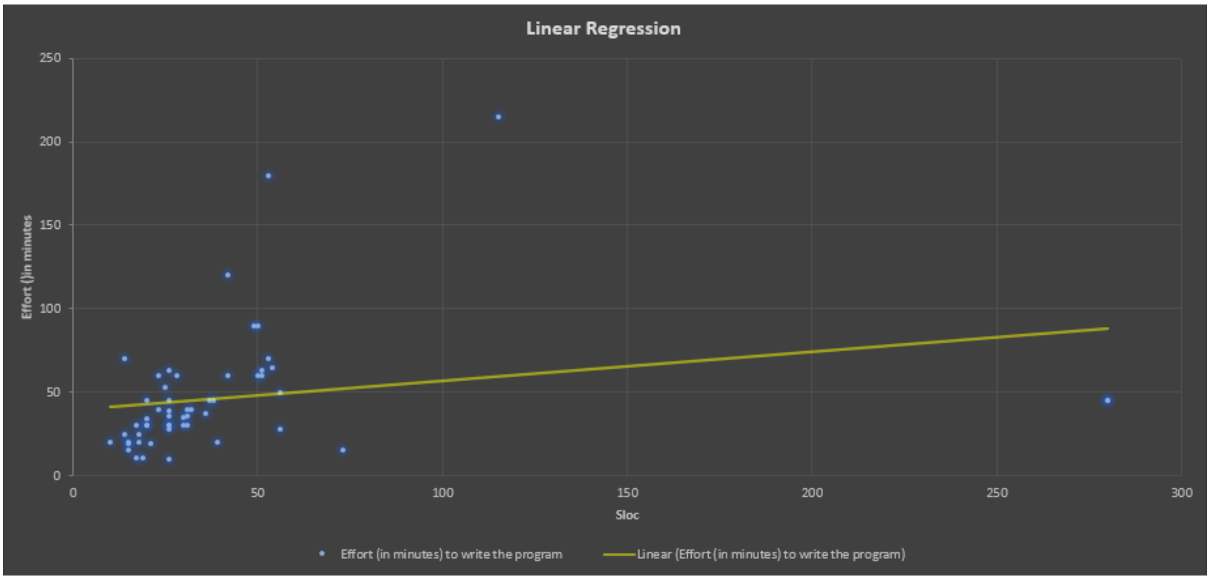
Σx_i^2 = 400+ 324+ 676+ 2809+ 289+ 361+ 225+ 400+ 1521+ 225+ 900+ 225+ 400+ 1764+ 676+ 676+ 2401+ 13225+ 1764+ 100+ 961+ 1369+ 2601+ 2601+ 324+ 784+ 5329+ 676+ 529+ 78400+ 529+ 676+ 676+ 676+ 3136+ 78400+ 900+ 196+ 3136+ 961+ 400+ 441+ 289+ 1296+ 2500+ 1444+ 196+ 961+ 676+ 1024+ 2916+ 625+ 2500+ 2809
$\Sigma x_i^2 = 230298$

$nx_{avg}^2 = 54 * (42.03703704)^2 = 95424.0741$

$\beta_1 = 129750.5 - (54 * 42.03703704 * 46.89814815) / (230298 – 95424.0741) = 23291.7975 / 134873.9259$
$\boldsymbol{\beta_1 = 0.1727}$

$\beta_0 = y_{avg} - \beta_1 * x_{avg}$
$\beta_0 = 46.89814815 - (0.1727 * 42.03703704)$
$\boldsymbol{\beta_0 = 39.6384}$

y* = 39.6384 + 0.1727*x
where y is the Effort in minutes and x is the length of code in SLOC



The above chart is attached in the submission (***A2-data-2022-solution.xslx***)

Via regression analysis, we optimally fitted a line to the datapoints and we can see that distance between the line and datapoints are minimized. The data points are close to the regression line except for the few outliers. The linear regression formula has been calculated on historical estimated and actual data for all the students.

2.2 Assumptions made in effort estimation model

Following assumptions were made in my effort estimation modelling:

- The assumption was made that the variables and data points are correct and relevant for estimation.
- The assumption was made that data is normally distributed with no data transformation applied on them.
- The assumption was made that outlier data will not affect analysis and the estimation model

Part 3: Validate empirically the prediction power of your estimation model

3.1 Firstly, estimate the effort for each SLOC in the TEST-A2-data

The regression line equation derived above is 39.6384 + 0.1727*X
The estimated effort below is in the excel sheet attached with the submission (***TEST-A2-data-solution.xslx***)

SLOC	Estimated effort	Work effort
31	44.9921	35 mins
21	43.2651	18 mins
32	45.1648	16 mins
16	42.4016	34 mins 43 secs
133	62.6075	191 mins
22	43.4378	1 hour 11 mins = 71 mins
12	41.7108	6:17 mins
15	42.2289	22 mins
29	44.6467	105 mins
38	46.201	15 mins

13	41.8835	37 mins
31	44.9921	52 minutes.
25	43.9559	20 minutes
10	41.3654	23 minutes
10	41.3654	15 minutes
73	52.2455	16 minutes and 47 seconds
55	49.1369	24 minutes
10	41.3654	22 minutes and 50 seconds
20	43.0924	63 minutes
16	42.4016	45 minutes
22	43.4378	35 Minutes
9	41.1927	21 minutes
19	42.9197	38 minutes
20	43.0924	10:47 minutes
9	41.1927	13 minutes 43 sec
47	47.7553	45 minutes
17	42.5743	60 minutes
12	41.7108	: 20 minutes
33	45.3375	40min
28	44.474	45 mins
29	44.6467	34 minutes
39	46.3737	41 mins 5 seconds
27	44.3013	38 mins
29	44.6467	237 mins
9	41.1927	27 mins
3	40.1565	13 min
10	41.3654	31 mins 12 seconds
23	43.6105	25 mins 48 seconds
15	42.2289	15 minutes 39 seconds
20	43.0924	43 mins
16	42.4016	28 mins
45	47.4099	60 minutes
13	41.8835	16 minutes
31	44.9921	60 mins
21	43.2651	41 mins
17	42.5743	26 minutes
15	42.2289	46 mins
35	45.6829	47 mins
25	43.9559	65 mins
10	41.3654	56 mins
31	44.9921	35 mins
27	44.3013	100 mins

3.2 Apply Coefficient of Determination R-square technique

Using the following formula:

Coefficient of Determination = MSS / TSS
where
TSS – Total Sum of Squares = Σ (Yi – Ym)2
MSS – Model Sum of Squares = Σ (Y^ – Ym)2
Y^ is the predicted effort value of using the effort estimation model,
Yi is the ith value of ith work effort reported by the programmer, and Ym is the mean value of all work effort data points reported by the programmers.

Ym = (35+ 18+ 16+ 34.7167+ 191+ 71+ 6.284+ 22+ 105+ 15+ 37+ 52+ 20+ 23+ 15+ 16.7834+ 24+ 22.83+ 63+ 45+ 35+ 21+ 38+ 10.7834+ 13.7167+ 45+ 60+ 20+ 40+ 45+ 34+ 41.084+ 38+ 237+ 27+ 13+ 31.2+ 25.8+ 15.65+ 43+ 28+ 60+ 16+ 60+ 41+ 26+ 46+ 47+ 65+ 56+ 35+ 100) / 52
Ym = 43.20861923

TSS = ∑(Yi - Ym)^2 = 67.38142966+ 635.4744835+ 740.3089604+ 72.11269221+ 21842.29223+ 772.3608451+ 1363.427505+ 449.8055296+ 3818.174737+ 795.7261989+ 38.54695274+ 77.28837584+ 538.6400066+ 408.3882912+ 795.7261989+ 698.2922114+ 368.9710527+ 415.2881217+ 391.6987528+ 3.209045063+ 67.38142966+ 493.2227681+ 27.12971428+ 1051.394842+ 869.7732999+ 3.209045063+ 281.9504682+ 538.6400066+ 10.29523736+ 3.209045063+ 84.79866812+ 4.514006872+ 27.12971428+ 37555.09926+ 262.7193373+ 912.5606758+ 144.2069358+ 303.0600235+ 759.4774939+ 0.043521983+ 231.3020989+ 281.9504682+ 740.3089604+ 281.9504682+ 4.877998903+ 296.1365758+ 7.791806603+ 14.37456814+ 474.8642759+ 163.619422+ 67.38142966+ 3225.26093
TSS = 83482.74812

MSS = ∑(Y^ - Ym)^2 = 3.180803657+ 0.003190077+ 3.826643205+ 0.651280038+ 376.3165751+ 0.052523825+ 2.243462446+ 0.95984977+ 2.068076301+ 8.954342673+ 1.755940974+ 3.180803657+ 0.558428549+ 3.39745713+ 3.39745713+ 81.66521405+ 35.14451289+ 3.39745713+ 0.013506909+ 0.651280038+ 0.052523825+ 4.063930342+ 0.083474321+ 0.013506909+ 4.063930342+ 20.67230602+ 0.402360886+ 2.243462446+ 4.532133333+ 1.601188493+ 2.068076301+ 10.01773628+ 1.193951265+ 2.068076301+ 4.063930342+ 9.315431794+ 3.39745713+ 0.161508153+ 0.95984977+ 0.013506909+ 0.651280038+ 17.65076011+ 1.755940974+ 3.180803657+ 0.003190077+ 0.402360886+ 0.95984977+ 6.122065329+ 0.558428549+ 3.39745713+ 3.180803657+ 1.193951265
MSS = 641.4640382

Coefficient of determination = MSS / TSS = 641.4640382/83482.74812 =
Coefficient of determination = 0.0076837916

Interpretation:

If r^2 is	the relationship is
$.9 \le r^2$	predictive; use it with high confidence
$.7 \le r^2 < .9$	strong and can be used for planning
$.5 \le r^2 < .7$	adequate for planning but use with caution
$r^2 < .5$	not reliable for planning purposes

The coefficient of determination is 0.007683 which less than 0.5. Hence it means the variation between MSS and TSS is too large and hence it is not reliable for planning purposes. Our estimation model is not reliable as it does not capture the relationship and the variation among variables correctly. Hence the data points need to be analyzed and the outliers should be dealt with in order to improve our model and achieve a better coefficient of determination.