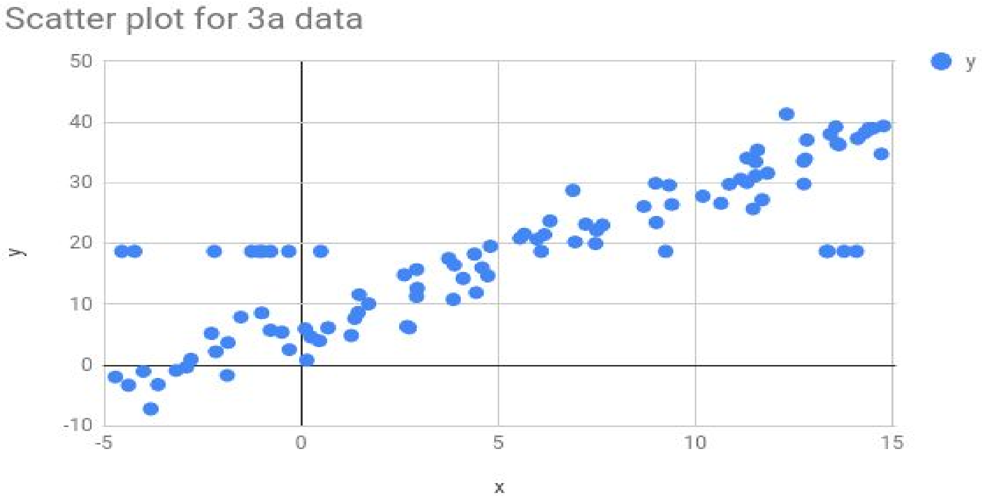
Report

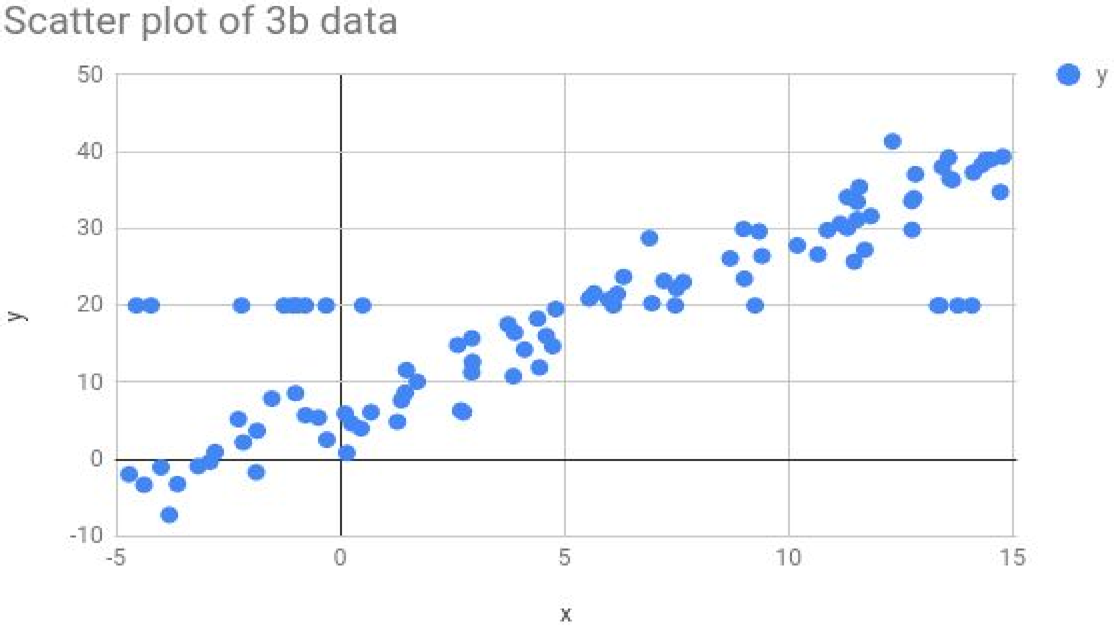
1) Both the variables are real. There are 15 data points missing out of 100 entries. Y is the variable dependent on variable X. Most of the entries are positive, hence mean and median should be positive.

2) No the missing data values cannot be ignored because the total number of data points are 100 out of which 15 are not available. On ignoring these points, the training data size will drastically reduce thus reducing the overall accuracy of the model.

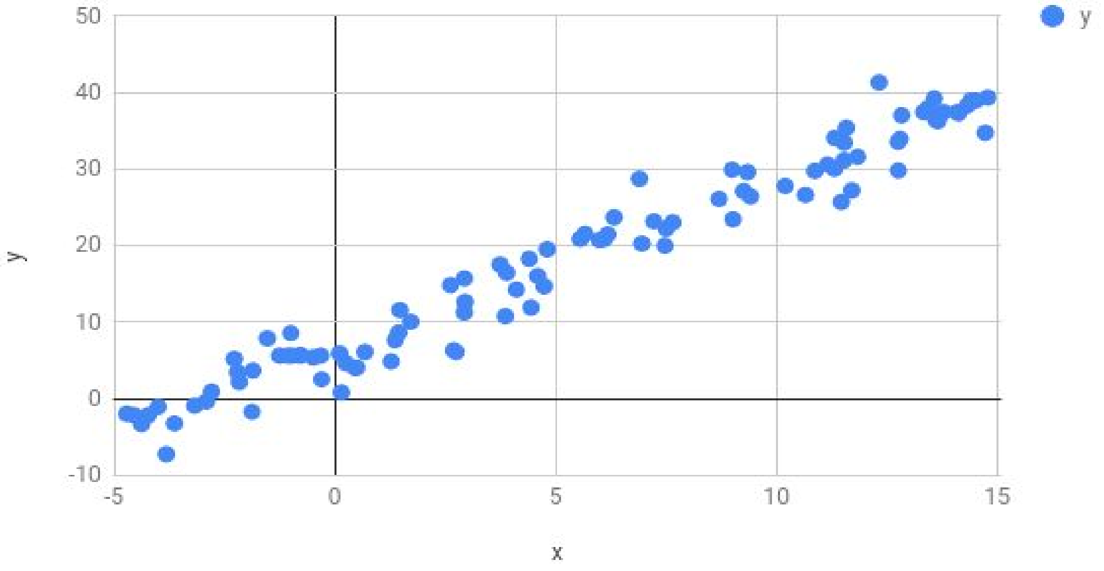
3) The created datasets are saved in the folder submitted.

4) (a) Scatter Plot of Y wrt X for each of the dataset

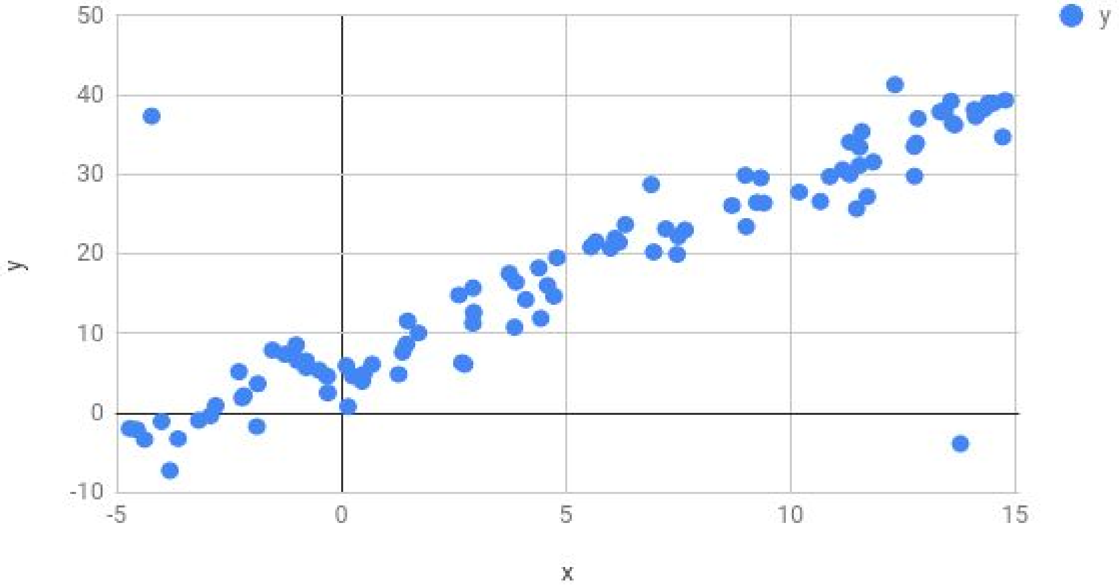


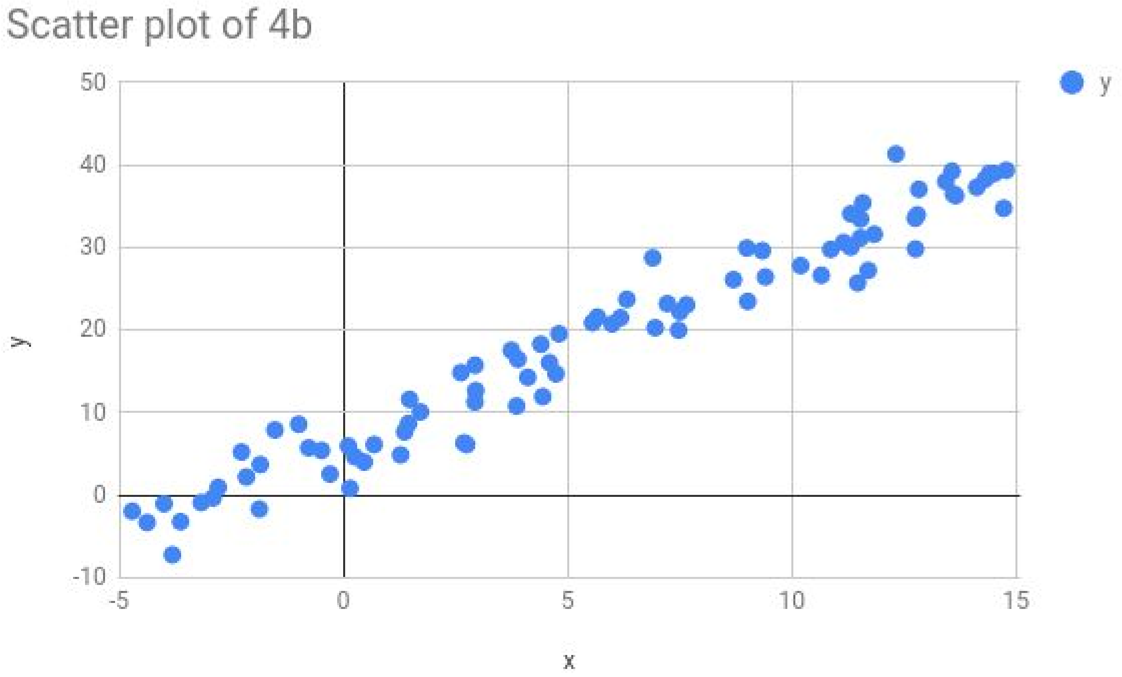


Scatter plot for 3c data



Scatter plot of 3d data





(b)

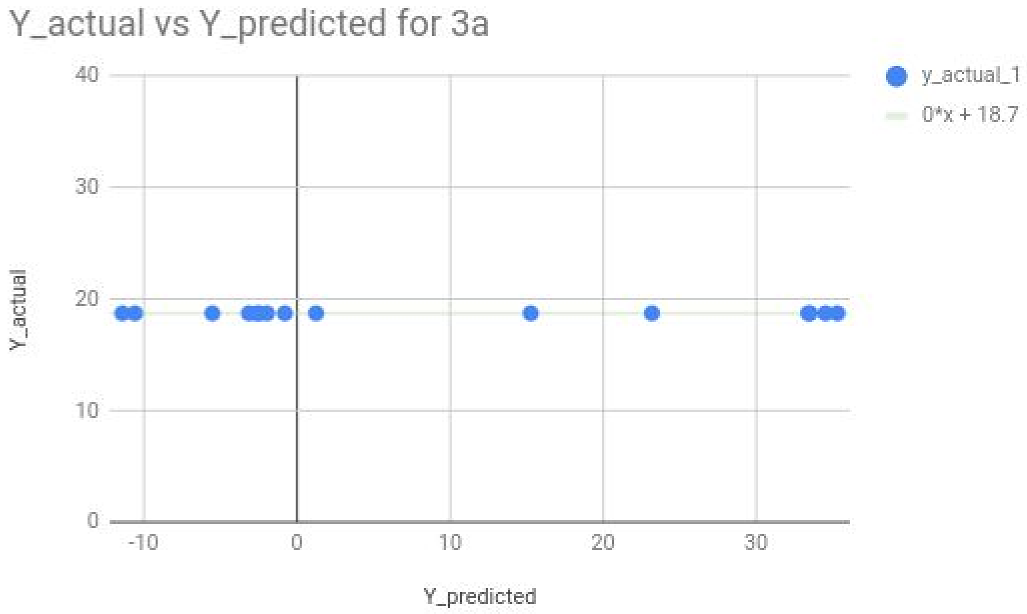
Case	R-Value	Adjusted R	p-value	F-Statistic	RMSE	MAE
3a	0.397	0.397	2.18e-32	inf	6.99	5.88
3b	0.475	0.475	1.86e-33	inf	12.09	8.96
3c	0.859	0.859	1.13e-52	inf	6.55	5.84
3d	0.564	0.564	3.35e-33	inf	4.25	3.58
4b	0.844	0.844	1.22e-39	inf	5.30	4.35

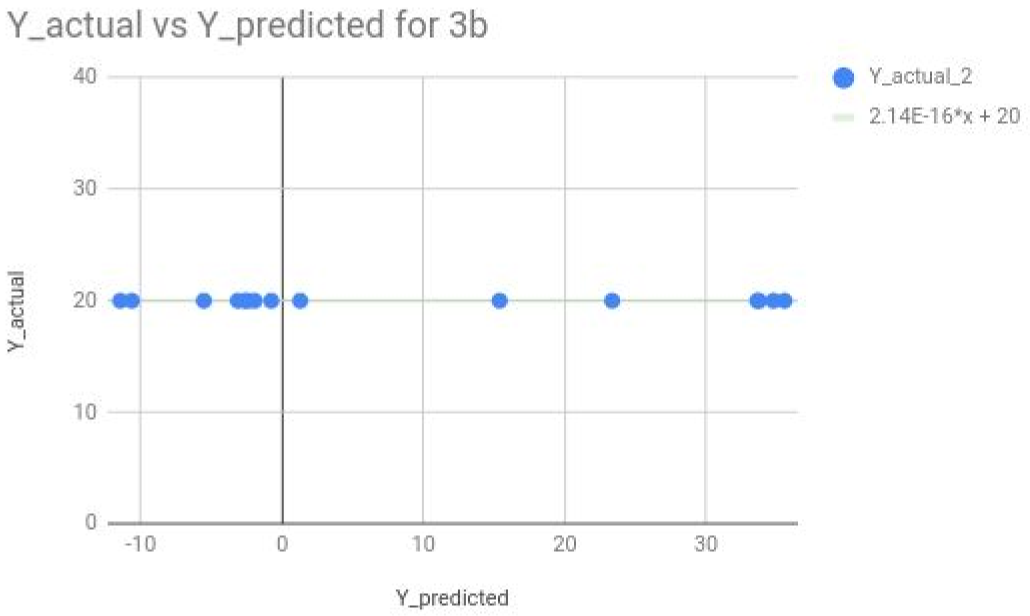
(c)

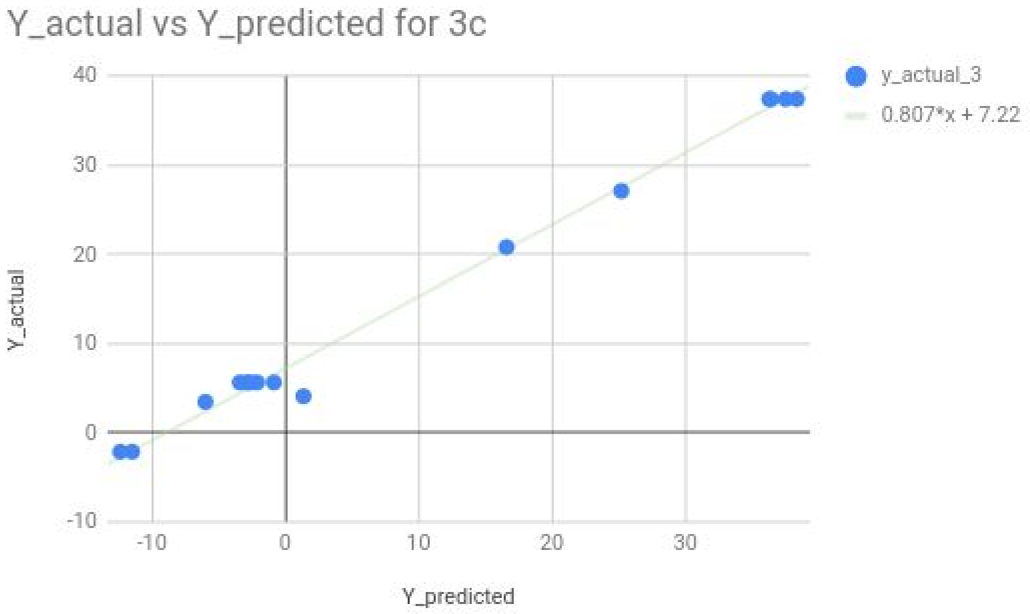
Case 3a	Lookin at R and Adjusted R value(0.397) suggests that, there is large variability in data which the model is not bae to capture. The p value in nearly 0 which means that null hypothesis of coefficient beig 0 is false, and the coefficient is meaningful addition to model. The RMSE and MAe values are also reasonably good, This suggests that the coefficient predicted are good but the data variability is very high.
Case 3b	Lookin at R and Adjusted R value(0.475) suggests that, there has been an improvement in variability prediction of data than the previous case.

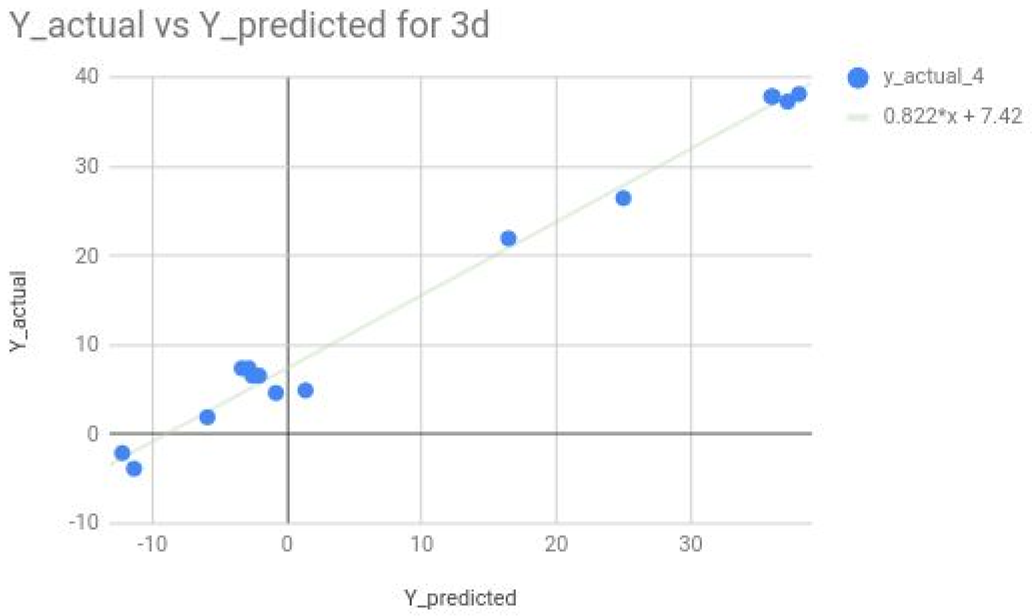
	The p value in nearly 0 which means that null hypothesis of coefficient beig 0 is false, and the coefficient is meaningful addition to model. There has been a reduction in data variability.
Case 3c	Lookin at R and Adjusted R value(0.564) suggests that, this model captures variability. The p value in nearly 0 which means that null hypothesis of coefficient beig 0 is false, and the coefficient is meaningful addition to model. Thus this suggests that using the Regression trees is one of the options for data imputations. The error values are too quiet, low suggesting high accuracy
Case 3d	Lookin at R and Adjusted R value(0.397) suggests that, this model is better that 3a and 3b but worse than 3c for data imputation. The p value in nearly 0 which means that null hypothesis of coefficient beig 0 is false, and the coefficient is meaningful addition to model.
Case 4b	Lookin at R and Adjusted R value(0.844) suggests that, there is large variability in data which the model is not bae to capture. The p value in nearly 0 which means that null hypothesis of coefficient beig 0 is false, and the coefficient is meaningful addition to model. But the RMSE is high suggesting less number of data points for training.

(d)
i) Plot of y-actual va y-residual



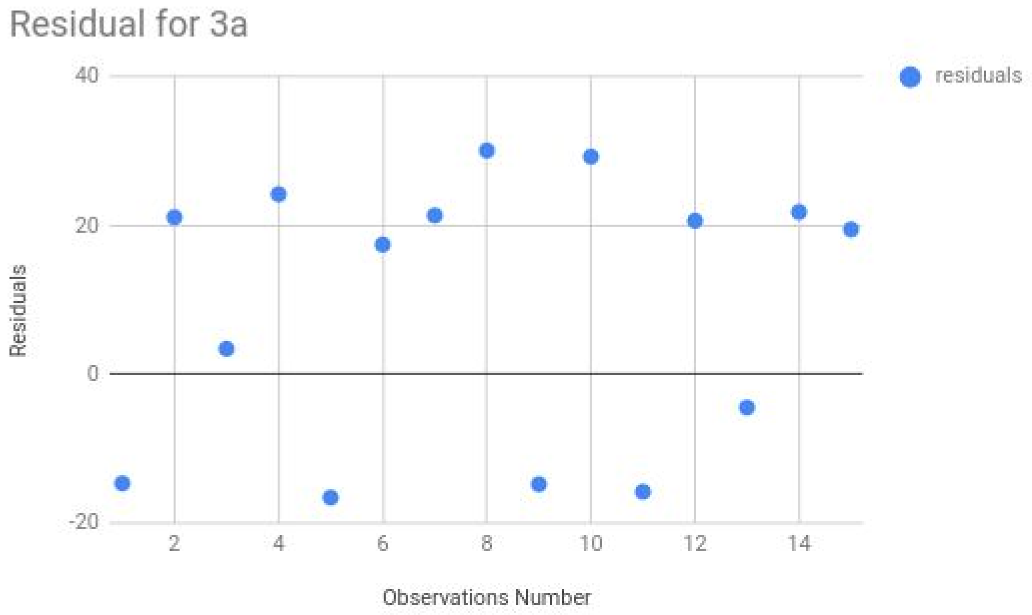

Y_actual vs Y_predicted for 3b
40
30
20
10
0
-10
0
10
20
30
Y_actual
Y_predicted
● Y_actual_2
2.14E-16*x + 20

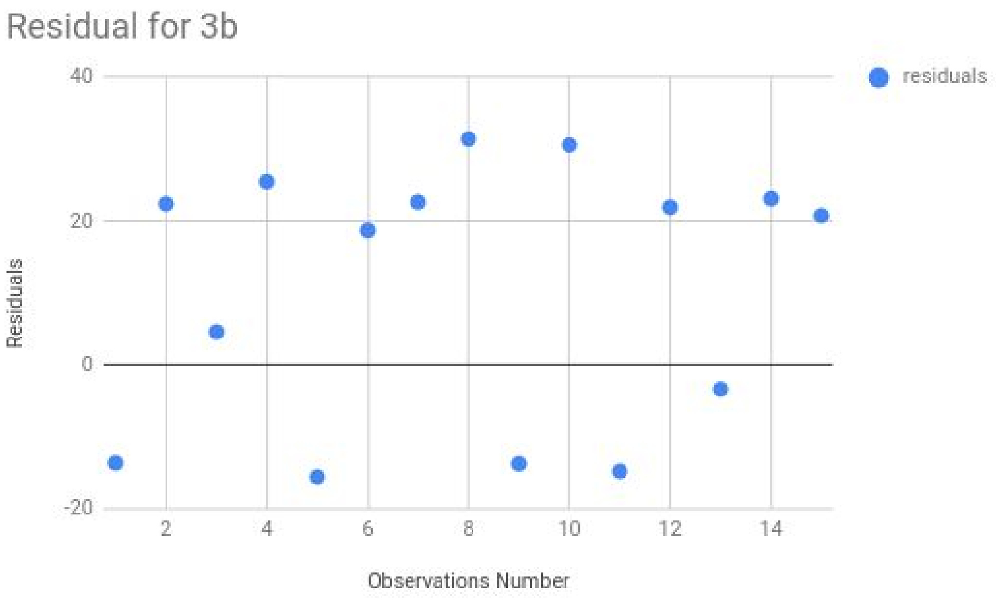

Y_actual vs Y_predicted for 3c
40
30
20
10
0
-10
-10
0
10
20
30
Y_actual
Y_predicted
● y_actual_3
0.807*x + 7.22

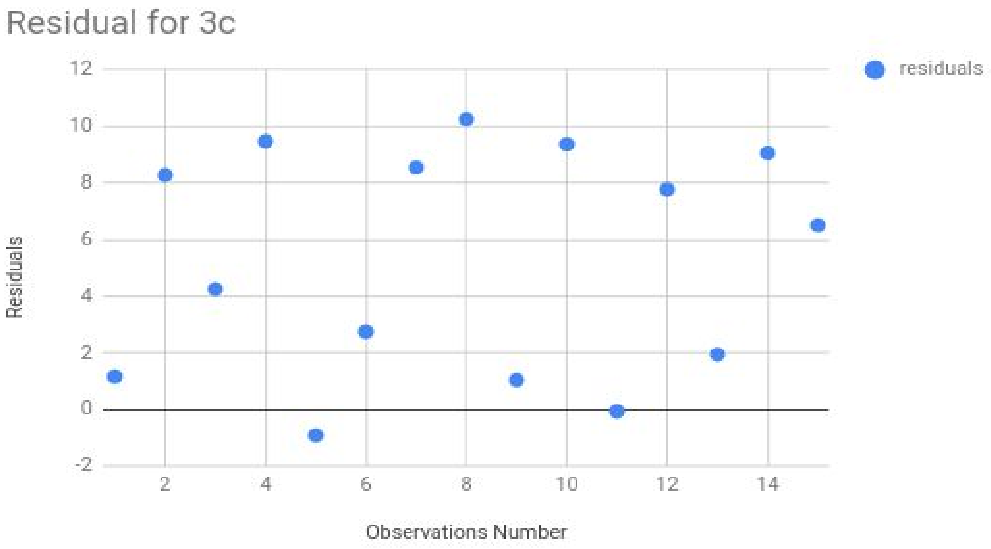


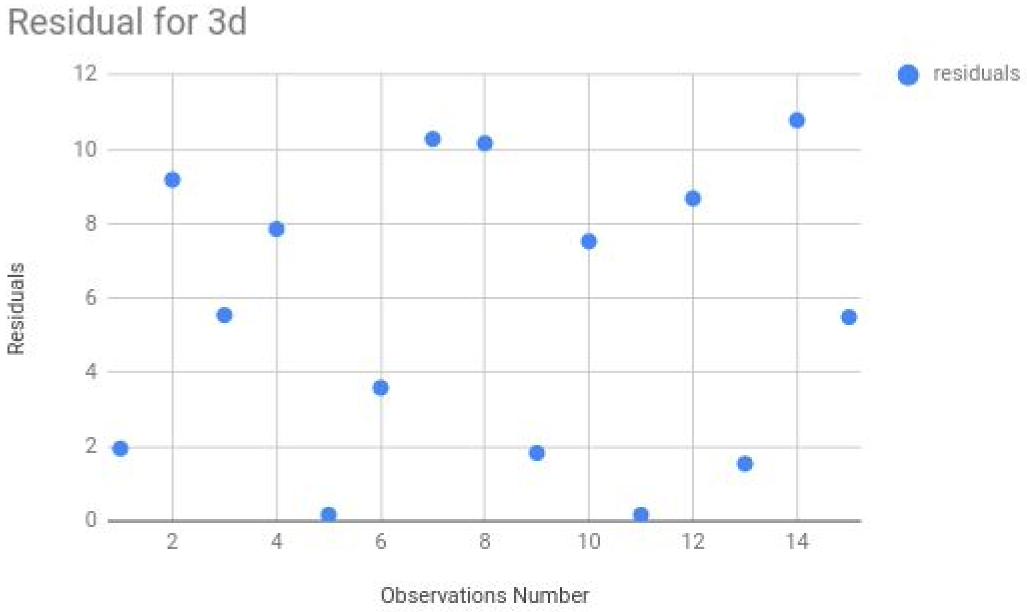
It can be seen from these graphs that the accuracy of 3c and 3d is more than that of 3a and 3b. It is visible form the equation of the trendline.

(ii) Residuals vs the Number of observation

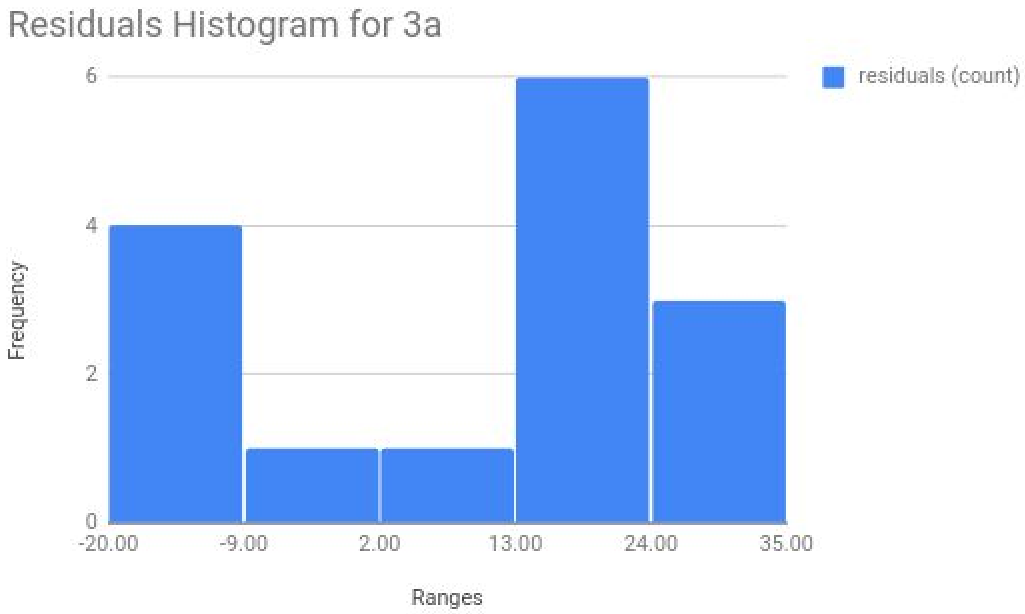


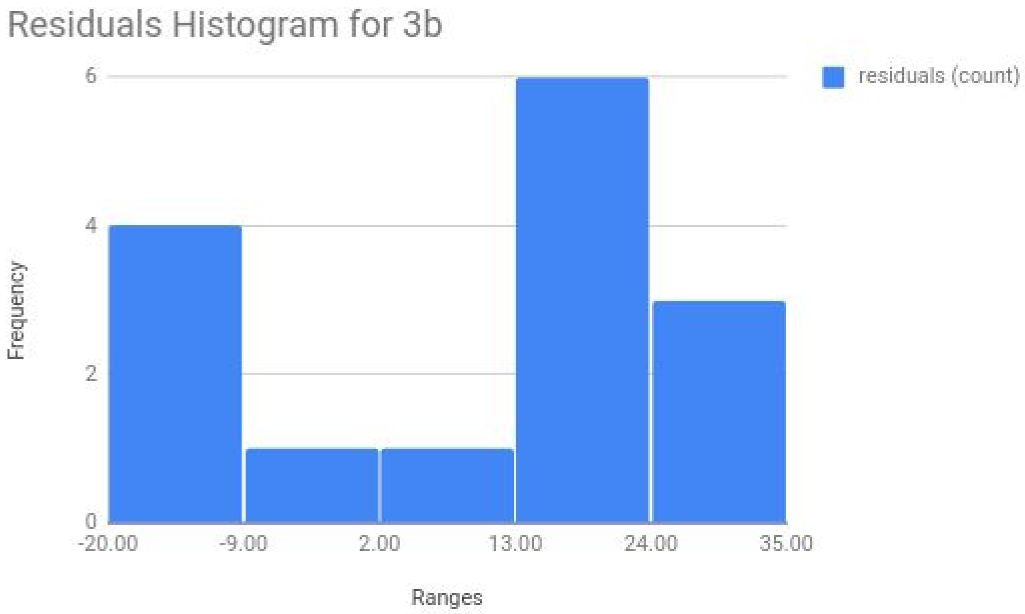

Residual for 3b
residuals
Residuals
40
20
0
-20
2
4
6
8
10
12
14
Observations Number

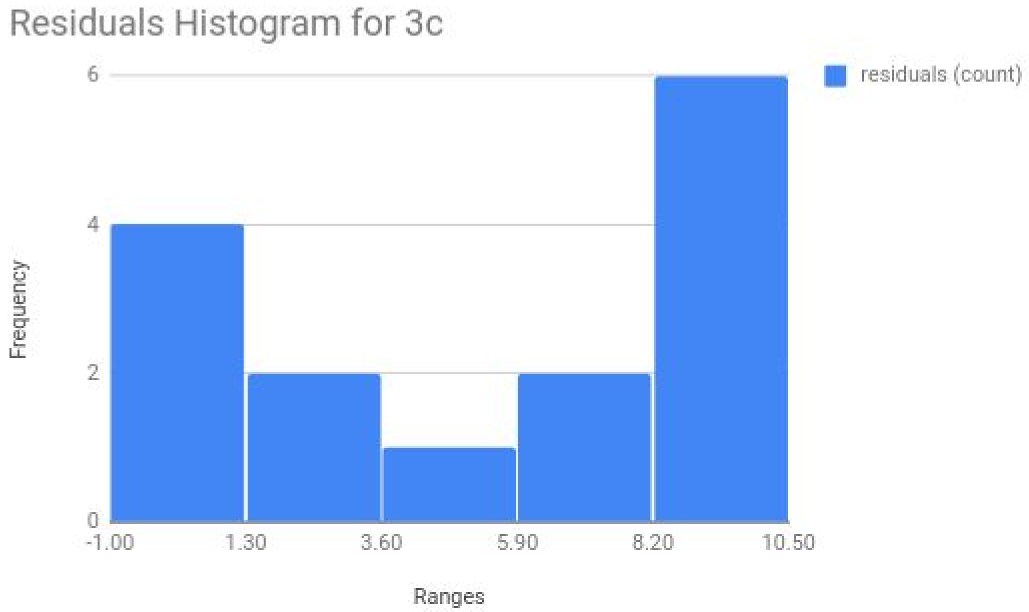

Residual for 3c
residuals
Residuals
12
10
8
6
4
2
0
-2
2
4
6
8
10
12
14
Observations Number

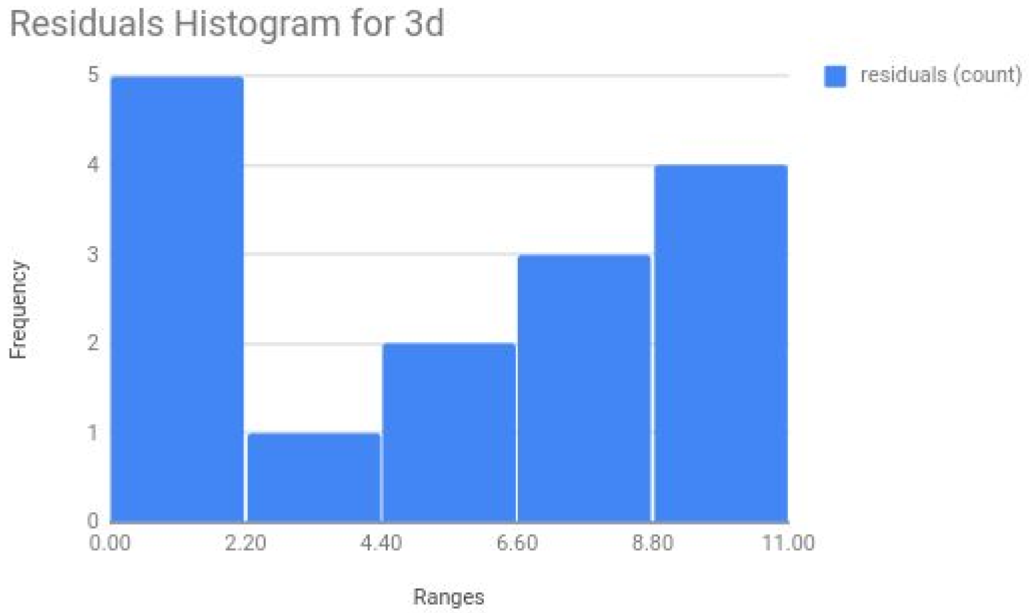

Residual for 3d
residuals
Residuals
12
10
8
6
4
2
0
2
4
6
8
10
12
14
Observations Number

(iii)Histogram of the Residuals-









Yes, the shape of histogram is almost same in all the cases.The shape of the histogram is like a pit in all the cases, with values minimum in between. This shape implies that very few of the predictions will have error of 0. Most of them will have some finite error.

Q 5)

The OLS model for 3c is the best to represent data. This is because on looking at the results in the table one can infer-

R-value and Adjusted R value=0.859, this is the highest among all which suggests this model captures most of the variation in the datasets.

P-value is almost 0, which significe the value of coefficients is very important and disproves the hypothesis of them being 0.
RMSE and MAE errors have lower value hence describes their accuracy.
Thus this concludes that regression trees are one of the best ways to impute data, Its accuracy can be further increased by using bagging and random forest.