Data Science Mini Project

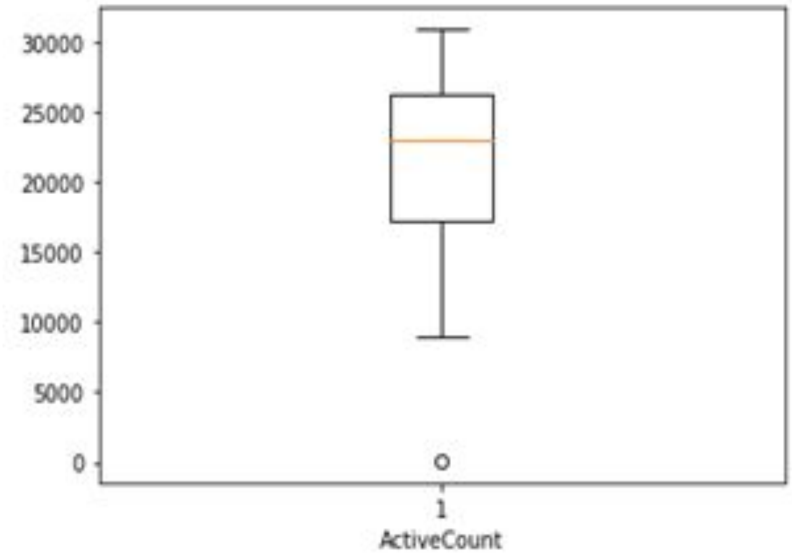
Group 7 (Tuesday Batch)

DATA :

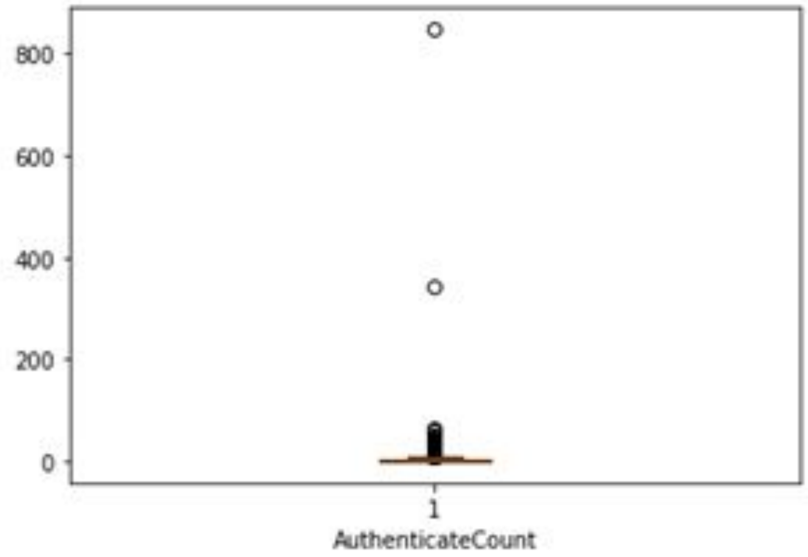
- Performance Record of a BNG device.
- Tracked at intervals of 15 min. on a day.
- CSV file: 11015 tuples of data, each with 14 attributes.(no missing values!)
- Train:Test Ratio is 70:30
- Problem Statement: To predict the **InBandWidth** using various regression techniques.

DESCRIPTIVE ANALYSIS

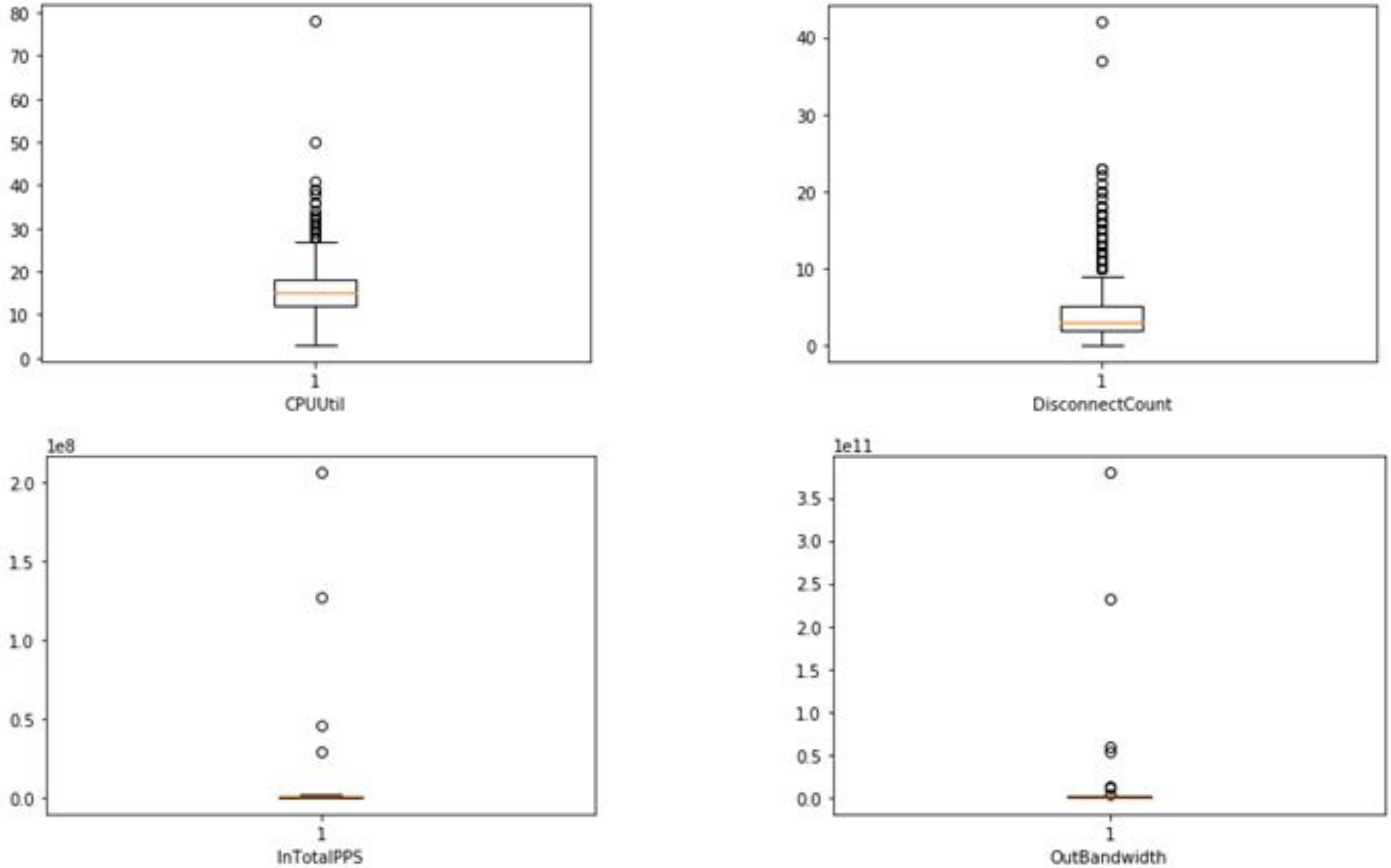
Boxplots -

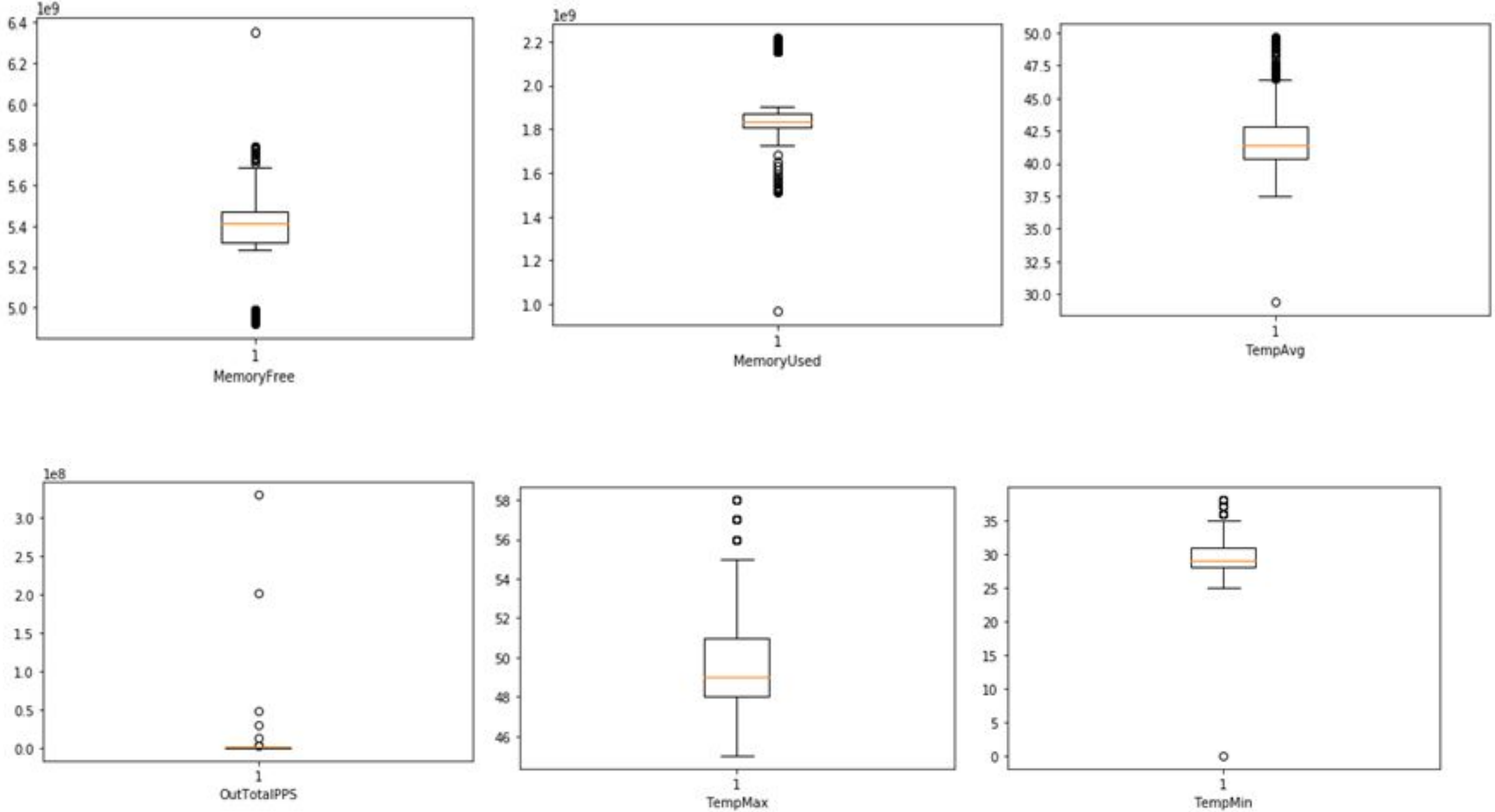


30000
25000
20000
15000
10000
5000
0
1
ActiveCount

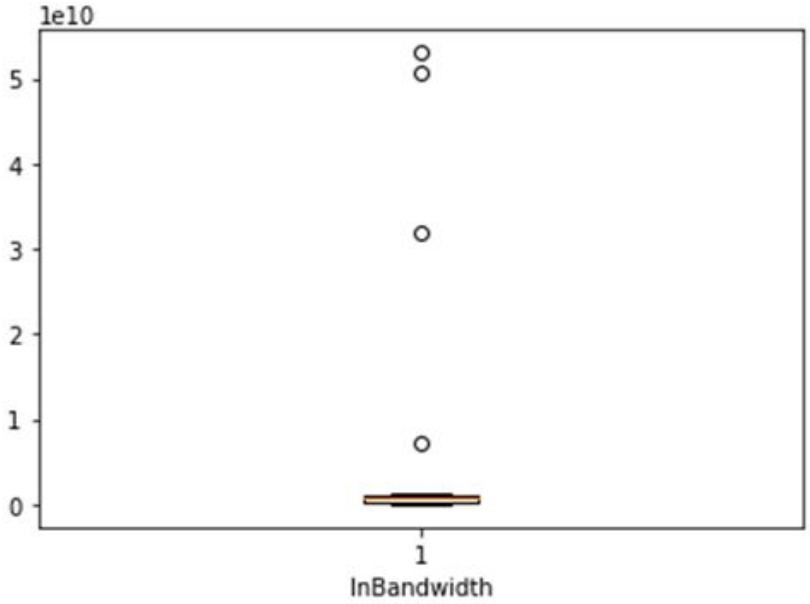


800
600
400
200
0
1
AuthenticateCount

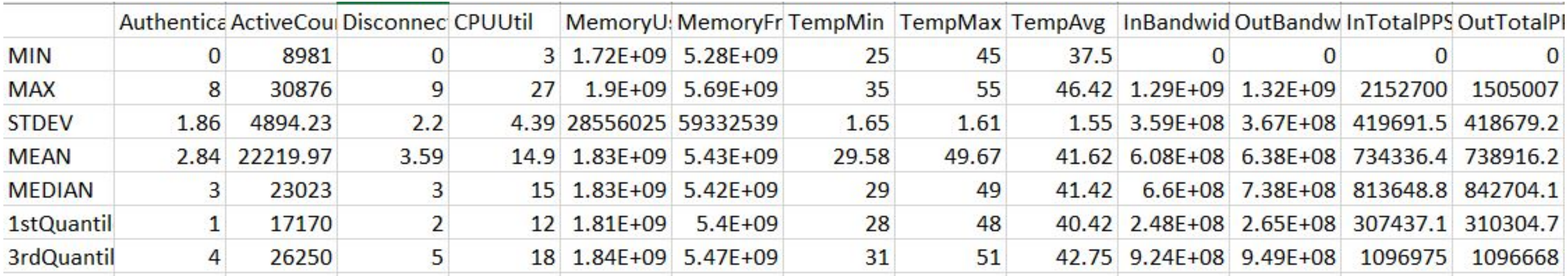

80
70
60
50
40
30
20
10
0
1
CPUUtil
40
30
20
10
0
1
DisconnectCount
1e8
2.0
1.5
1.0
0.5
0.0
1
InTotalPPS
1e11
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
OutBandwidth


1e9
6.4
6.2
6.0
5.8
5.6
5.4
5.2
5.0
1
MemoryFree
1e9
2.2
2.0
1.8
1.6
1.4
1.2
1.0
1
MemoryUsed
50.0
47.5
45.0
42.5
40.0
37.5
35.0
32.5
30.0
1
TempAvg
1e8
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
OutTotalPPS
58
56
54
52
50
48
46
1
TempMax
35
30
25
20
15
10
5
0
1
TempMin

Boxplot (Target Atrribute)



STATS OF DATA :



	Authentica	ActiveCou	Disconnec	CPUUtil	MemoryU	MemoryFr	TempMin	TempMax	TempAvg	InBandwid	OutBandw	InTotalPPS	OutTotalPI
MIN	0	8981	0	3	1.72E+09	5.28E+09	25	45	37.5	0	0	0	0
MAX	8	30876	9	27	1.9E+09	5.69E+09	35	55	46.42	1.29E+09	1.32E+09	2152700	1505007
STDEV	1.86	4894.23	2.2	4.39	28556025	59332539	1.65	1.61	1.55	3.59E+08	3.67E+08	419691.5	418679.2
MEAN	2.84	22219.97	3.59	14.9	1.83E+09	5.43E+09	29.58	49.67	41.62	6.08E+08	6.38E+08	734336.4	738916.2
MEDIAN	3	23023	3	15	1.83E+09	5.42E+09	29	49	41.42	6.6E+08	7.38E+08	813648.8	842704.1
1stQuantil	1	17170	2	12	1.81E+09	5.4E+09	28	48	40.42	2.48E+08	2.65E+08	307437.1	310304.7
3rdQuantil	4	26250	5	18	1.84E+09	5.47E+09	31	51	42.75	9.24E+08	9.49E+08	1096975	1096668

Descriptive Analysis

Outliers :

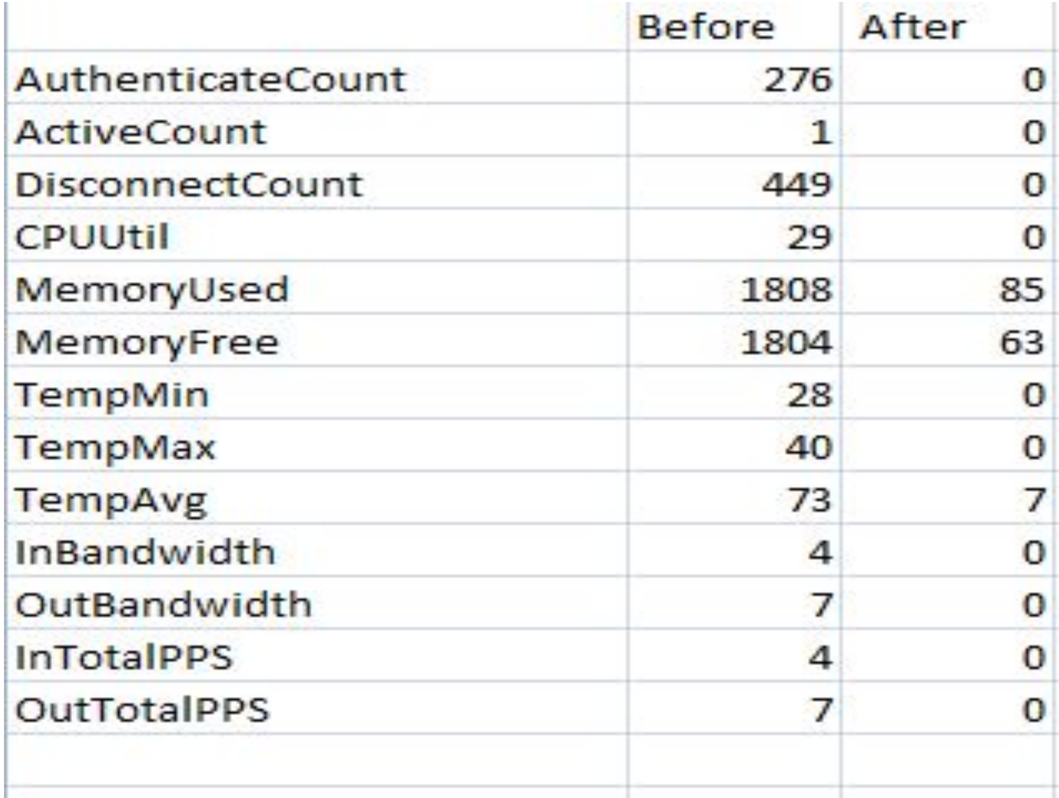
Total outliers before replacing with median = 4530

Total outliers after replacing with median = 155

Outliers in the Target Attribute:

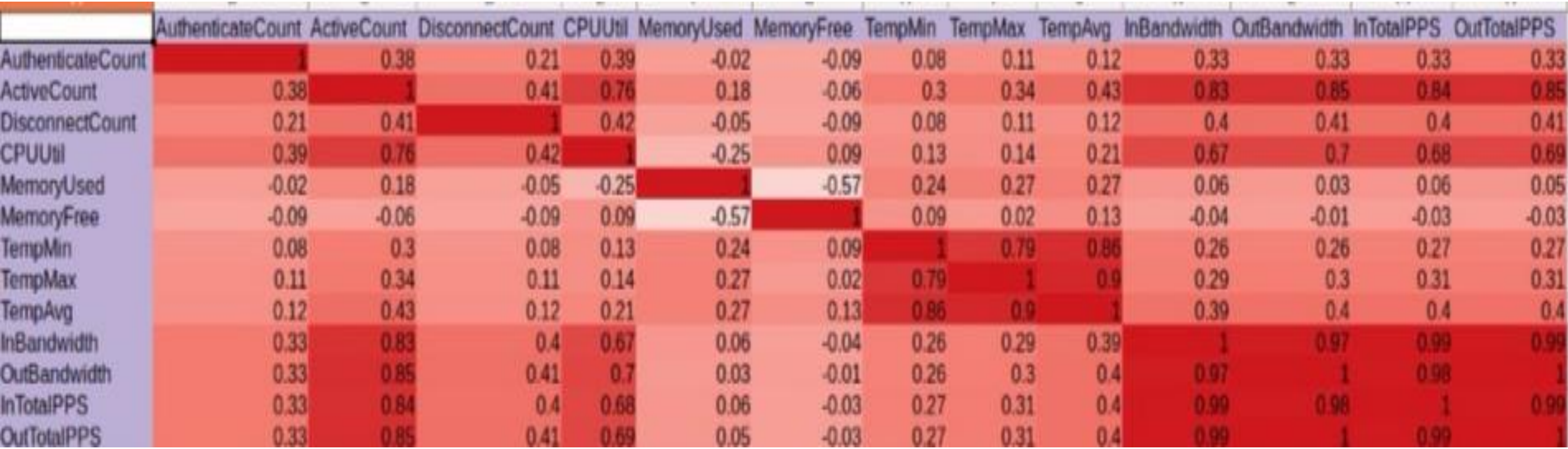
Before removal: 4 After removal: 0

Outliers Count -



	Before	After
AuthenticateCount	276	0
ActiveCount	1	0
DisconnectCount	449	0
CPUUtil	29	0
MemoryUsed	1808	85
MemoryFree	1804	63
TempMin	28	0
TempMax	40	0
TempAvg	73	7
InBandwidth	4	0
OutBandwidth	7	0
InTotalPPS	4	0
OutTotalPPS	7	0

Correlation matrix -



	AuthenticateCount	ActiveCount	DisconnectCount	CPUUtil	MemoryUsed	MemoryFree	TempMin	TempMax	TempAvg	InBandwidth	OutBandwidth	InTotalPPS	OutTotalPPS
AuthenticateCount	1	0.38	0.21	0.39	-0.02	-0.09	0.08	0.11	0.12	0.33	0.33	0.33	0.33
ActiveCount	0.38	1	0.41	0.76	0.18	-0.06	0.3	0.34	0.43	0.83	0.85	0.84	0.85
DisconnectCount	0.21	0.41	1	0.42	-0.05	-0.09	0.08	0.11	0.12	0.4	0.41	0.4	0.41
CPUUtil	0.39	0.76	0.42	1	-0.25	0.09	0.13	0.14	0.21	0.67	0.7	0.68	0.69
MemoryUsed	-0.02	0.18	-0.05	-0.25	1	-0.57	0.24	0.27	0.27	0.06	0.03	0.06	0.05
MemoryFree	-0.09	-0.06	-0.09	0.09	-0.57	1	0.09	0.02	0.13	-0.04	-0.01	-0.03	-0.03
TempMin	0.08	0.3	0.08	0.13	0.24	0.09	1	0.79	0.86	0.26	0.26	0.27	0.27
TempMax	0.11	0.34	0.11	0.14	0.27	0.02	0.79	1	0.9	0.29	0.3	0.31	0.31
TempAvg	0.12	0.43	0.12	0.21	0.27	0.13	0.86	0.9	1	0.39	0.4	0.4	0.4
InBandwidth	0.33	0.83	0.4	0.67	0.06	-0.04	0.26	0.29	0.39	1	0.97	0.99	0.99
OutBandwidth	0.33	0.85	0.41	0.7	0.03	-0.01	0.26	0.3	0.4	0.97	1	0.98	1
InTotalPPS	0.33	0.84	0.4	0.68	0.06	-0.03	0.27	0.31	0.4	0.99	0.98	1	0.99
OutTotalPPS	0.33	0.85	0.41	0.69	0.05	-0.03	0.27	0.31	0.4	0.99	1	0.99	1

DATA PRE-PROCESSING

- Checking for missing or NaN values -> absent
- Checking for outliers -> replacement(median).
- Feature Selection by pearson correlation.
- PCA for dimensionality reduction.

PREDICTIVE ANALYSIS

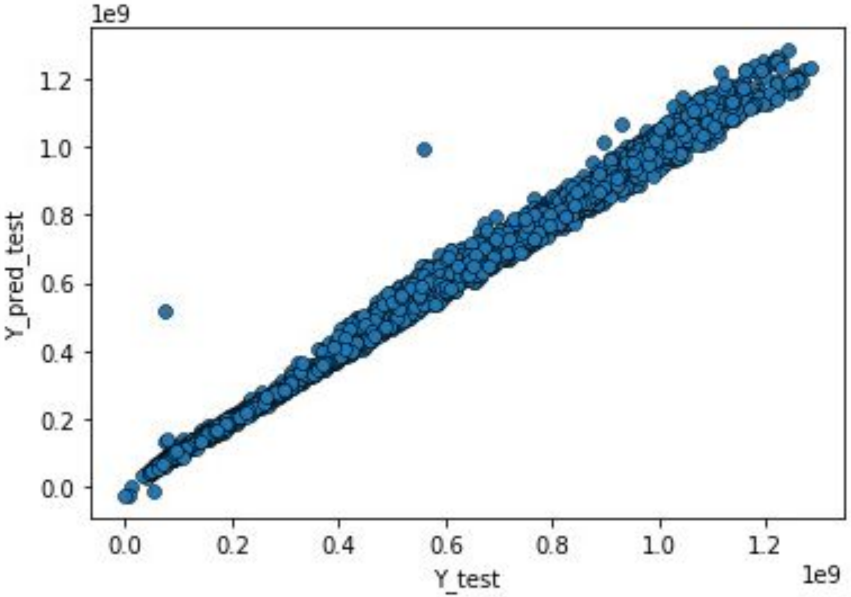
Linear Regression on Attributes :

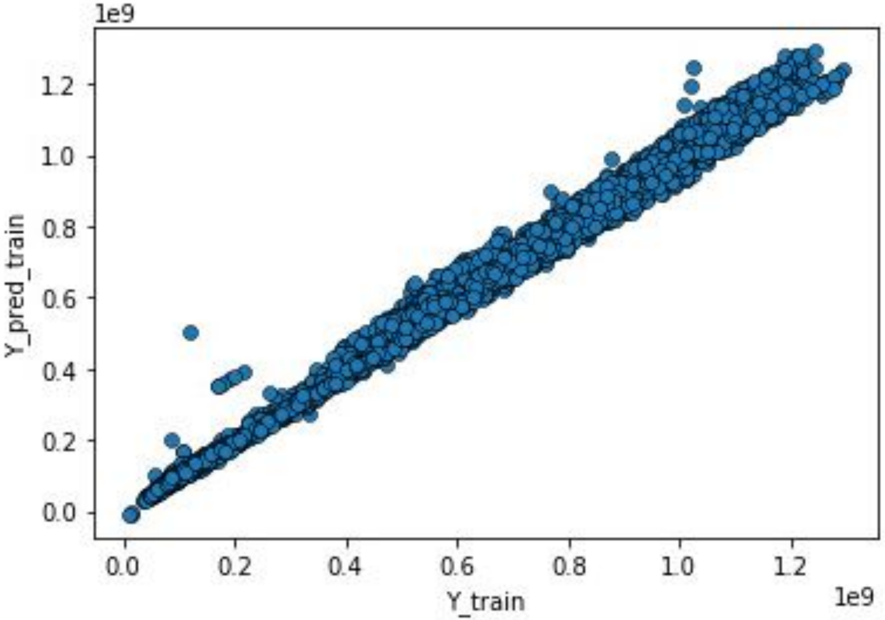
- **Attributes - Outbandwidth , InTotalPPS , OutTotalPPS**
- Correlation coefficient with InBandwidth > 0.96
- RMSE(train data) : 27323755.23
- RMSE(test data) : 28569495.24
- R2 score on train data = 0.9942
- R2 score on test data = 0.9935
- Model parameter :

  [-1.23244706e+00, 2.79709739e+02, 1.64456220e+03]
  W0=-25910825.924729705

Linear Regression on Attributes :

- Attributes - Outbandwidth , InTotalPPS , OutTotalPPS





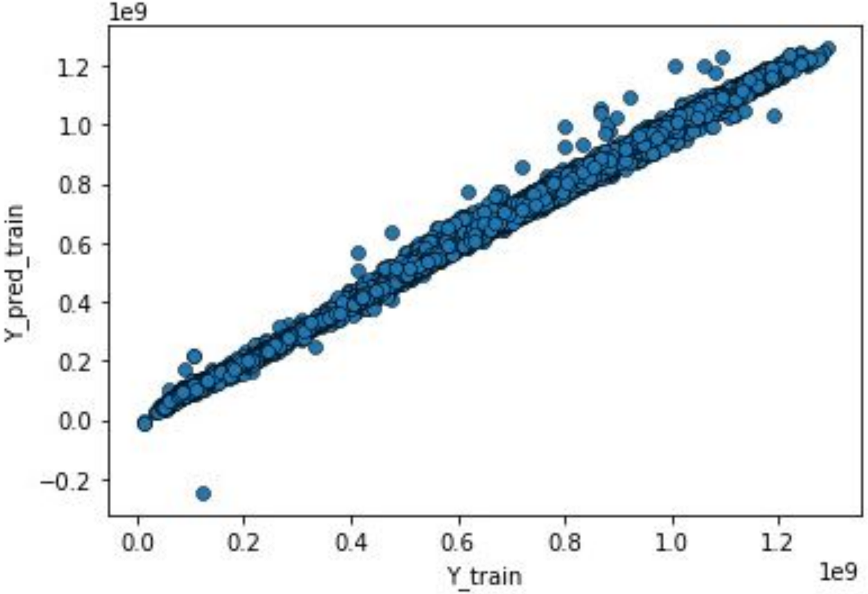
Polynomial Regression on Attributes :

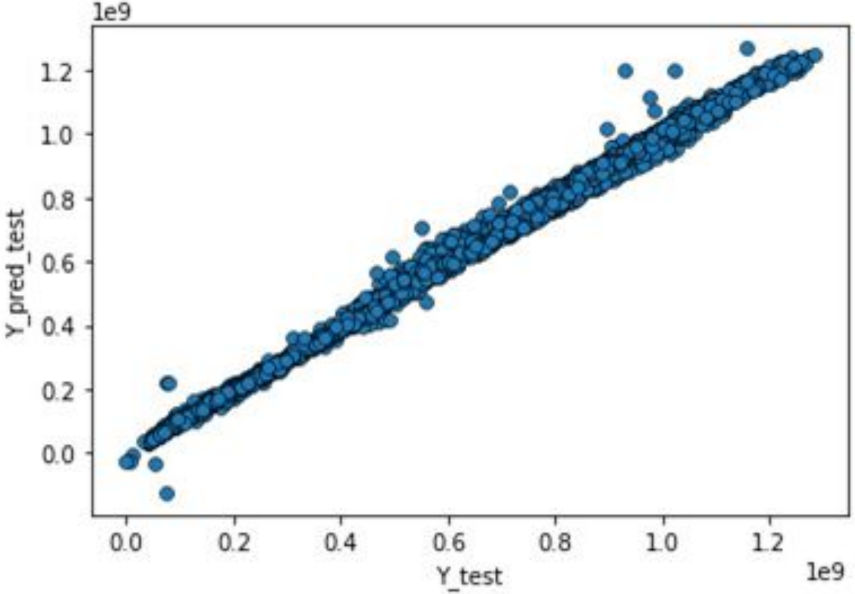
- **Attributes - Outbandwidth , InTotalPPS , OutTotalPPS**

- Correlation coefficient with InBandwidth > 0.96
- Optimal degree is 2
- RMSE(train data) : 19925569.9583
- RMSE(test data) : 20443378.0531
- R2 score on train data = 0.9969
- R2 score on test data = 0.9967
- Model parameter :
  [0.00000000e+00 , -1.56770074e+00 , 7.64346615e+02 , 1.47217372e+03 , -1.28228803e-08 , -1.01847998e-05 , 3.34430201e-05 , -9.35535833e-04 , 1.08879238e-02 , -2.04867721e-02]

Polymomial Regression on Attributes :

- Attributes - Outbandwidth , InTotalPPS , OutTotalPPS



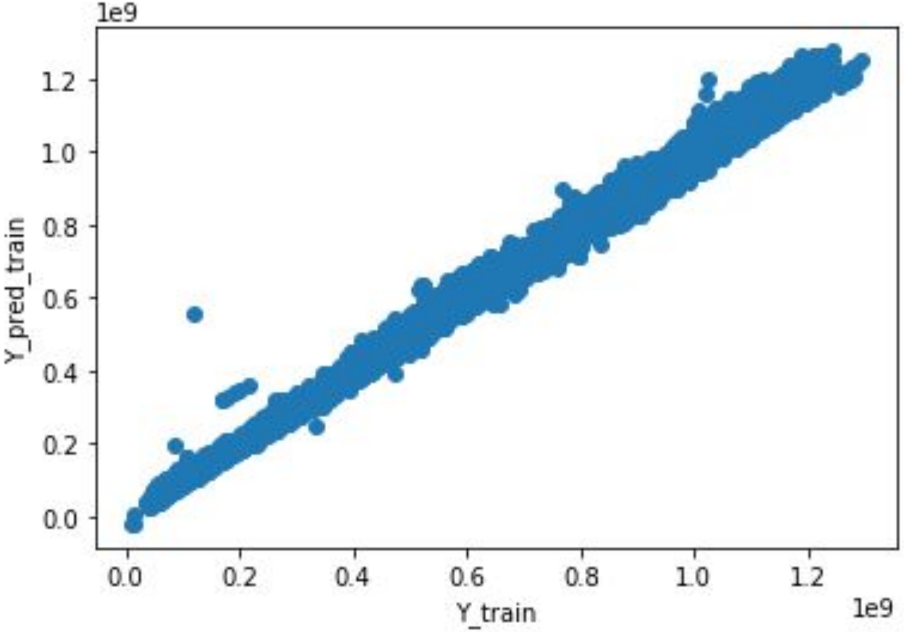


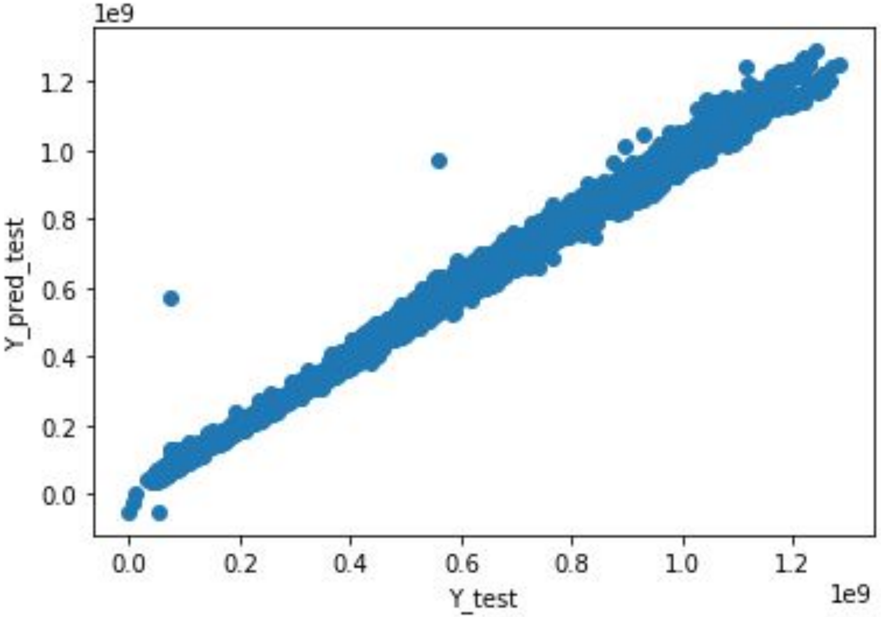
Linear Regression on PCA with optimal Attributes :

- **Optimal attributes are for Dimension 6**
- RMSE(train data) : 24673326.077913806
- RMSE(test data) : 26327007.115445804
- R2 score on train data = 0.9952916408358703
- R2 score on test data=   0.9945525724504367
- Model parameters:

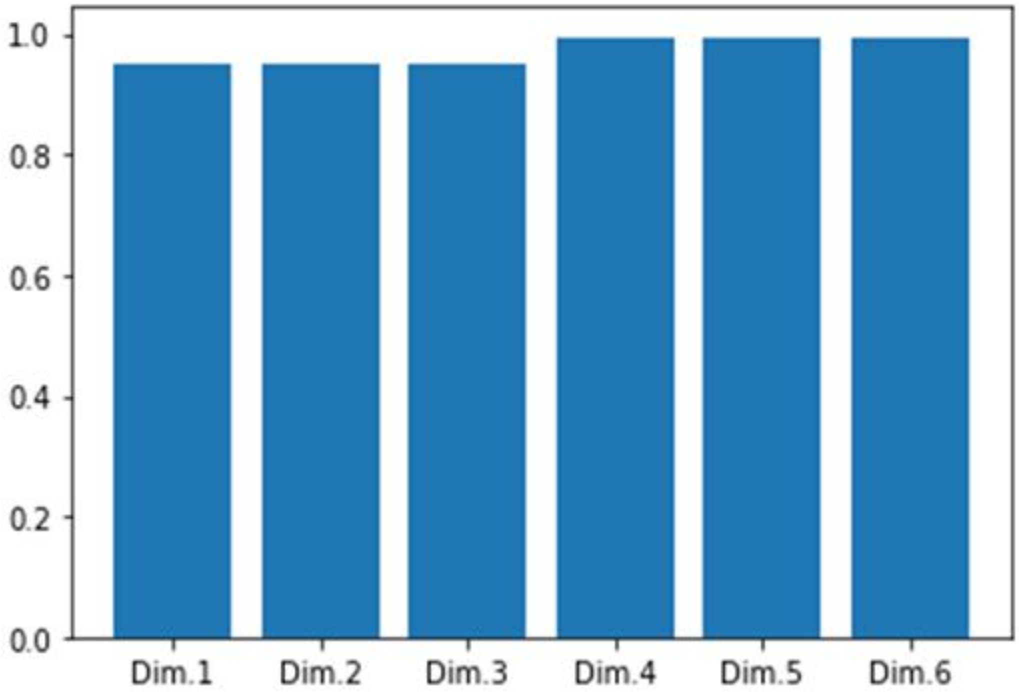
[-1.68612005e+03, 9.53003734e-01, -1.68386284e-01,
3.86217428e-01,9.06955769e+02, -1.49896543e+03 ,3.47369196e+03]

Linear Regression on PCA with optimal Attributes :





R2 - Score For Different Dimensions :



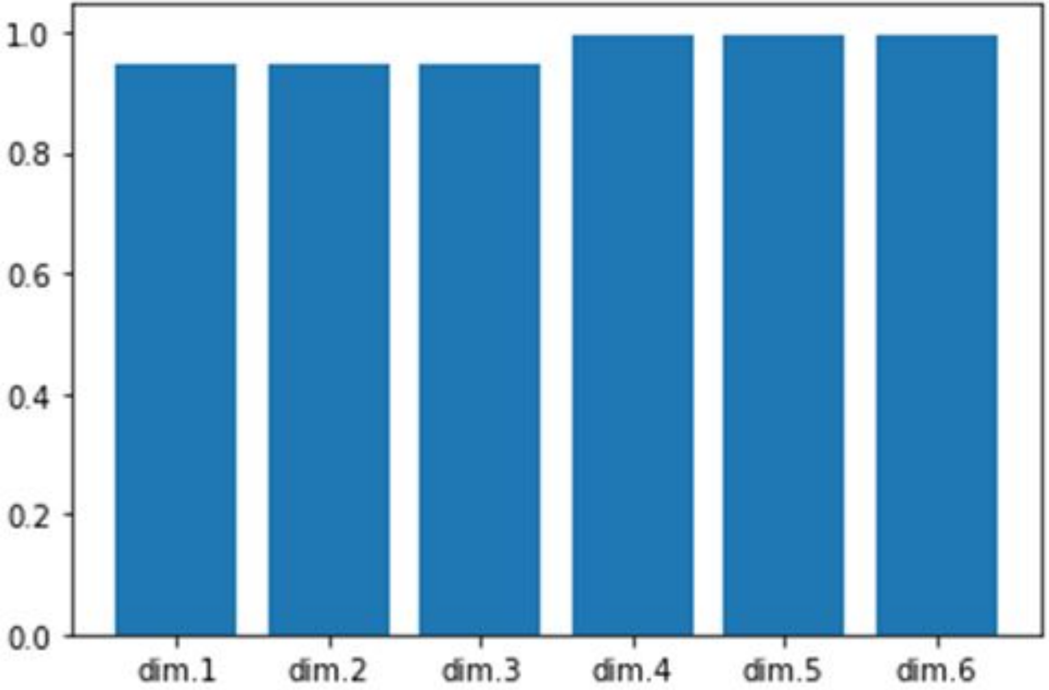
Inferences From Linear Regression

- Best results are obtained after applying PCA with components = 6
- After dimension=3 R2 score value saturates to 0.994.

Polynomial Regression on PCA with optimal Attributes :

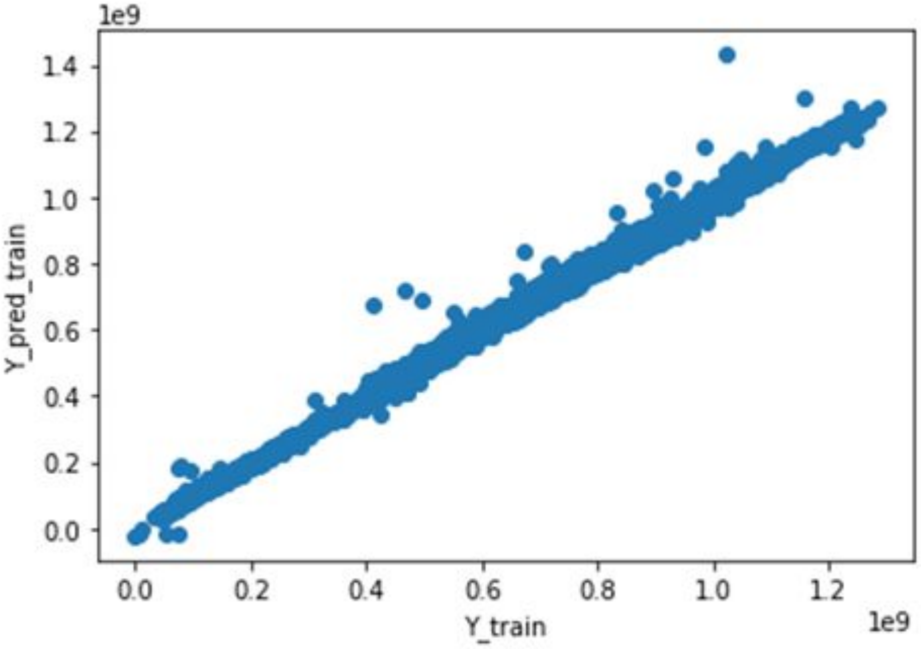
- **Optimal no. of Dimension = 5**
- RMSE(train data) : 16552484.048265139
- RMSE(test data) :19261364.398282487
- R2 score on train data = 0.9978808395334157
- R2 score on test data= 0.997088658767089

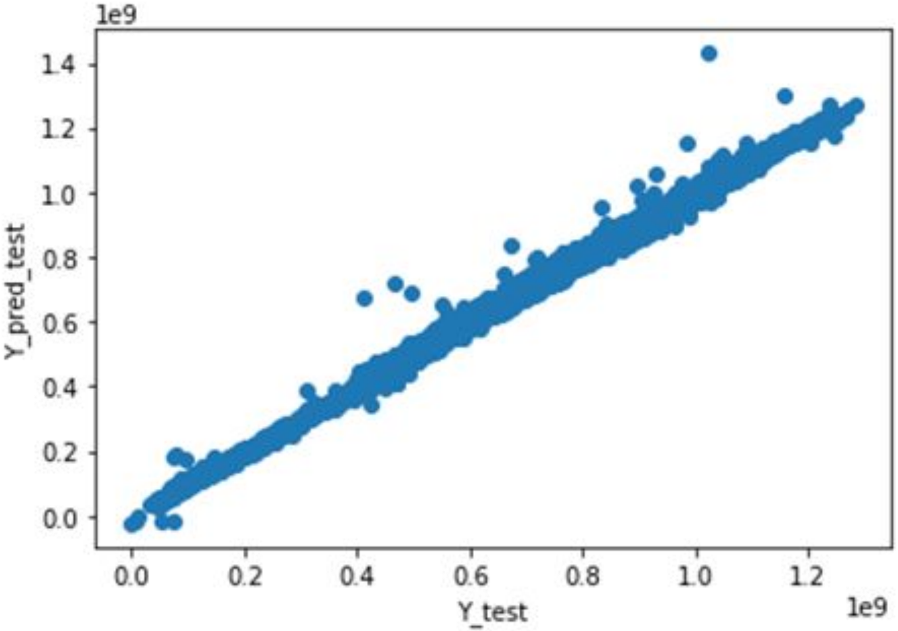
R2 - Score For Different Dimensions :



Degree=2

Polynomial Regression on PCA with optimal Attributes :

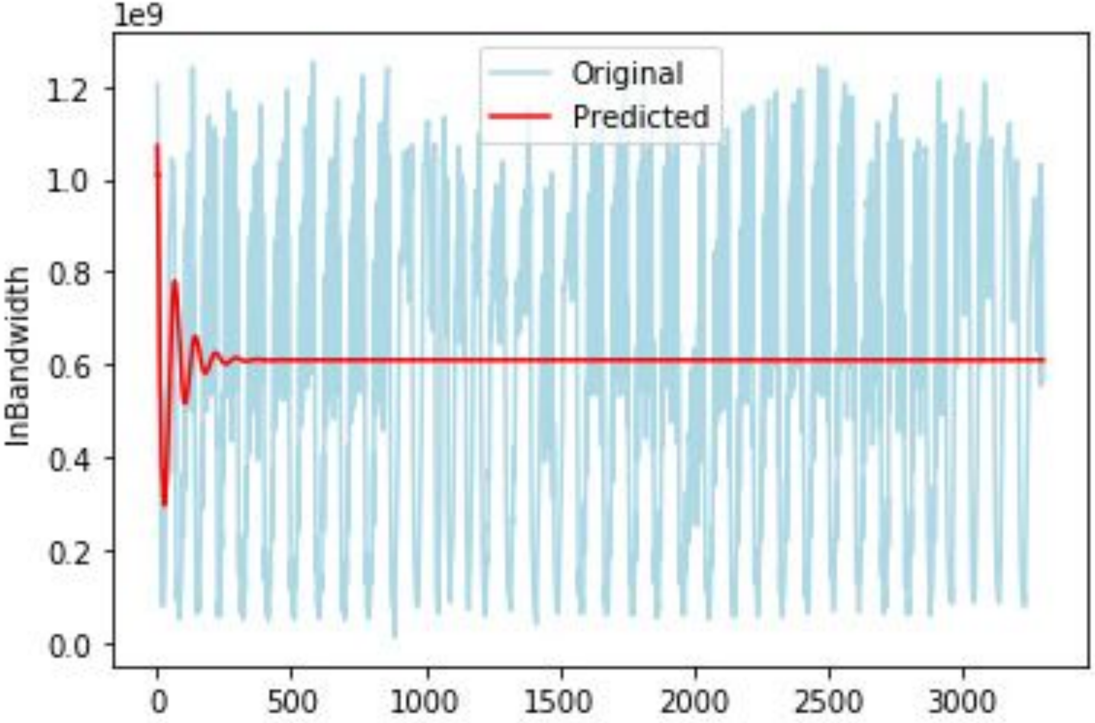




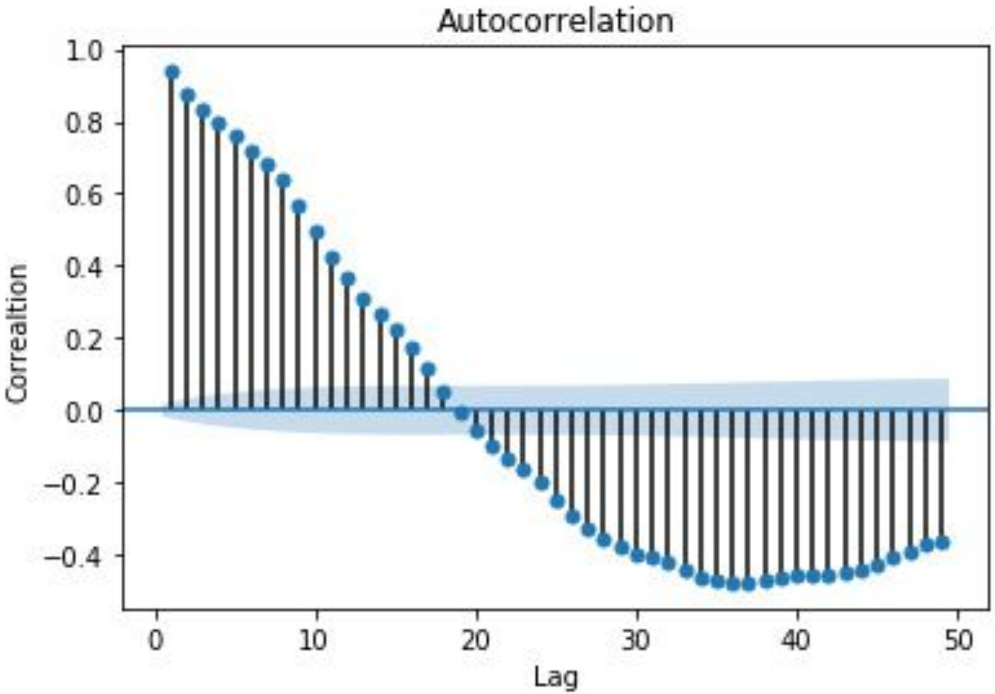
Inferences From Polynomial Regression

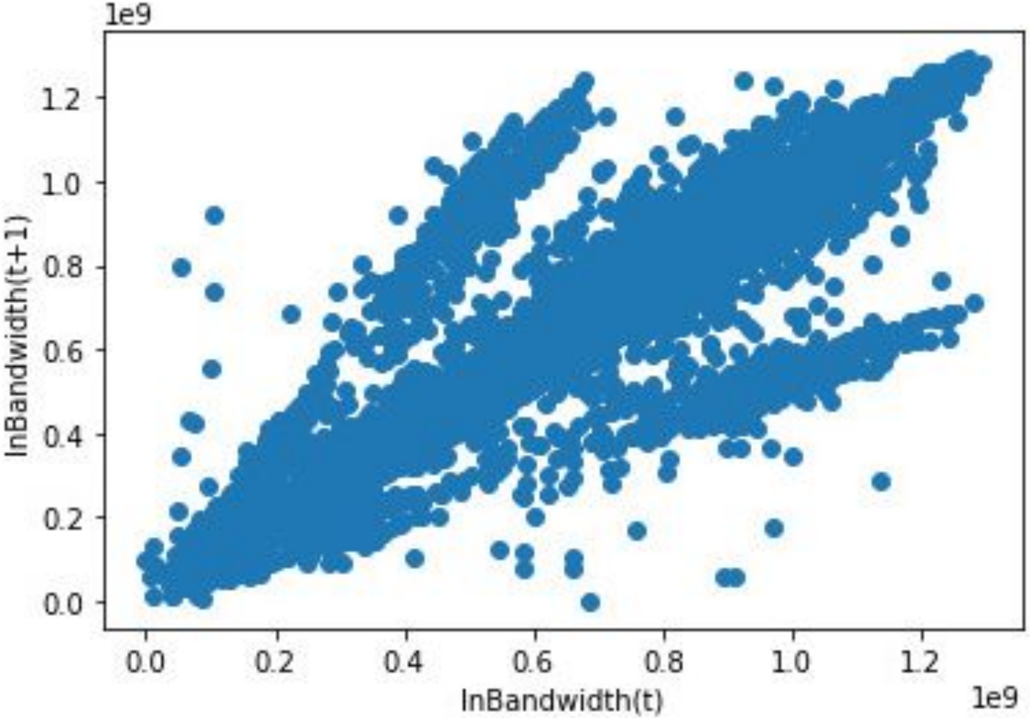
- Best Results are obtained when dimensions are reduced to 5
- The r2 score value saturates to 0.997 after dimension 5, on both train and test data.
- The degree of equation is 2 , a quadratic curve is fitting for the data.

AUTO - REGRESSION



Optimal Lag: 17


Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Correaltion
0
10
20
30
40
50
Lag


1e9
1.2
1.0
0.8
0.6
0.4
0.2
0.0
InBandwidth(t+1)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
InBandwidth(t)
1e9

Inferences From Auto Regression

- Auto Correlation
- Finding the optimal Lag value.
- Choosing lag = 1.

CONCLUSION

BEST MODEL

Polynomial regression of degree 2 with no. of dimension 5