Data Quality

Error Repairing Methods

Repairing Typo Error

1) Using build-in python library “pattern”
2) Build dictionary with the most common movies genre
3) Drop uncorrected words

Error in 25% of the data

	Dirty Data	Cleaned Data		
	MSE	**MSE - “pattern lib”**	**MSE - Dict of genre**	**MSE - Drop**
Lasso Without HP	1.28	0.89	0.89	0.91
Lasso With HP	1.12	0.85	0.85	0.87

Error in 50% of the data

	Dirty Data	Cleaned Data		
	MSE	**MSE - “pattern lib”**	**MSE - Dict of genre**	**MSE - Drop**
Lasso Without HP	1.30	0.87	0.91	0.91
Lasso With HP	1.13	0.84	0.87	0.86

Error in 75% of the data

	Dirty Data	Cleaned Data		
	MSE	**MSE - “pattern lib”**	**MSE - Dict of genre**	**MSE - Drop**
Lasso Without HP	1.32	0.89	0.89	0.89
Lasso With HP	1.16	0.86	0.88	0.86

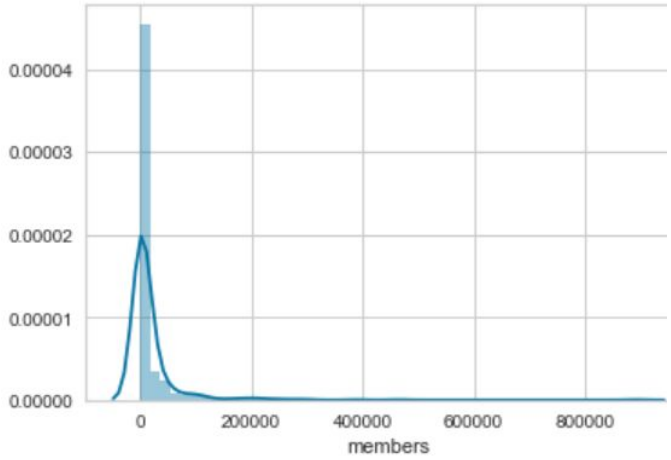
Repairing Gaussian Noise

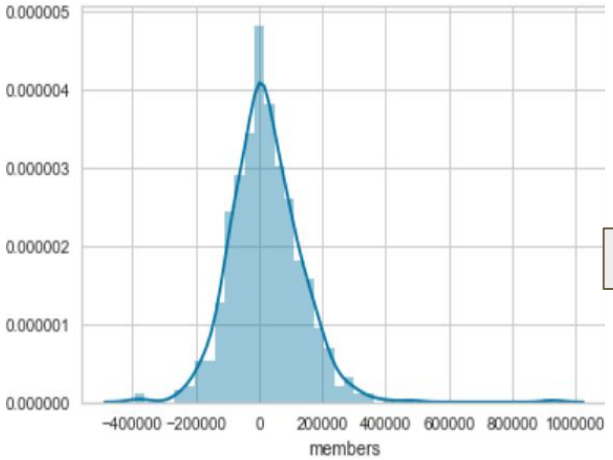
1) Using savgol_filter to filter the data from the noise
2) Dropping the negative values
3) Replacing negative values with 0
4) Get the absolute values of negative ones

Error in 30% of the data

	Dirty Data	Cleaned Data			
	MSE	**MSE - Filter**	**MSE - Abs**	**MSE - Drop**	**MSE - Replace**
Lasso Without HP	0.68	0.97	0.98	0.93	0.93
Lasso With HP	0.68	0.97	0.97	0.94	0.93

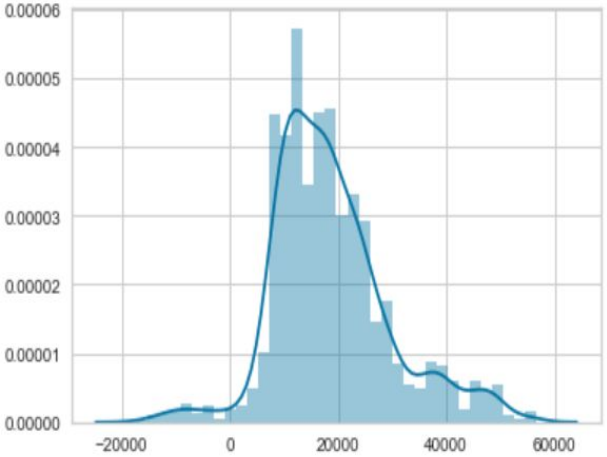
Original



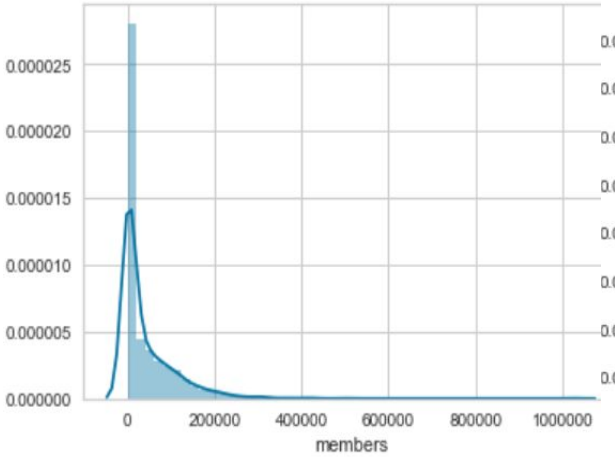


With Gaussian Noise

Filter



Cast



Drop / Replace