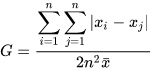
**Hybrid Gini Index:**

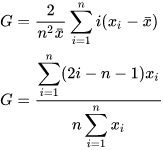
Gini Index is calculated based on the lorenz curve of income distribution.

The gini index equation based on rmse is as follows:



As per the points to ignore/eliminate negative values and zeros from the income data.

In reference to statdirect if the income data is in ascending order, and the values are positive then the gini index equation can be reduced to the below equation:



Based on the above equation, the income data is sorted in ascending order.

Then to ignore zero(0), add a least possible number ( in this case 10^-24) to it, such that it satisfies the above equation and doesn’t cause any significant error.

And to eliminate negative values shift the origin, that is, we have to move the origin to the least value in the set. Such that the range between the least value and the maximum value remains the same. And the Lorenz graph can be plotted with starting point from (0, 0) instead of (negative income(x), 0).

As the curve remains congruent even after shifting the origin, and the area between this curve the slope of 1 line (ideal income vs population line) will also remain the same

Thus, based on the observed congruency and least significant value added to zeros in the data results in identical gini values as that of rmse gini value.

Now, to decrease the computation taken to calculate for complex and large dataset. The best possible way is to reduce the data such that it can yield a similar gini index value.

As per the gini index rule, if the population’s income is closer to the mean, then the value remains closer to 1 or ideal situation and doesn’t affect the gini index value.

The anomaly only occurs when the lesser income value or higher income values are in place. The proposition of these either ends of the spectrum of the income data decides the index value.

So, it is ideal to keep these values intact as they are the most significant values in determination of the gini index for the data.

Thus, for any distribution the population in the interquartile range is closer to mean and the 1st and last quartile consist of least/highest values.

So, the possible solution is to keep the q1 and q4 as the same and extract a percentage of values from q2 and q3.

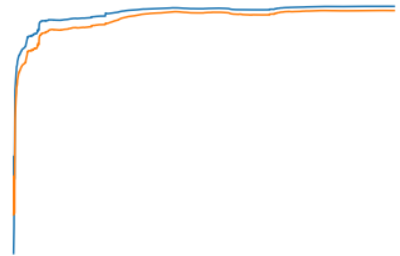
So, in the proposed solution, it extracts a random sample from selectable percentage of data from q2 and q3 and merges that with the existing q1 and q4 data.

Thus, the total number of data reduces and reduces the computation.

The observed difference from hybrid gini and actual gini is as follows:

Only for the first three weeks the difference is between 0.045 to 0.02

Later on the difference is between 0.015 and after a year its close to 0.005.



Graph between hybrid vs actual.