Reducing Inequalities using Big Data Techniques

Anmol Shukla	Bishal Kumar Shrestha	Parth Joshi	Jay Kakkad
112551470	112689393	112675594	112961206

1. Introduction

Inequality as quoted by the United Nations of Economic and Social Council (ECOSOC) is " social, economic and political challenge and has a profound impact on sustainable development". Today we face an existent issue of income inequalities, which has become more significant over the last decade. Income inequality as stated by the United Nations is the income gap between rich and poor. In 2018, it was observed that the top 20% of the population have earned 52% of all U.S. income, whereas the richest of the rich, the top 5% have earned 23% of all income. It is well known that the income gap between top 5 income earners and the rest, is the largest it has been since the last decade.

United Nations SDG #10 is to reduce this inequality who have set their target for 2030, where they aim to progressively achieve & sustain income growth of the bottom 40 percent of the population at a rate higher than the national average. The aim of this report is to analyze the hypothesis of existing income inequality, its extent and finding compelling correlation between sex, gender, cast, education and their effect on 4 income groups[1].

2. Background

After doing some background research about this topic, we found out that there are two main targets along with their indicators that are set by the United Nations Development Program.

Targets:

- By 2030, progressively achieve and sustain income growth of the bottom 40 percent of the population at a rate higher than the national average
- By 2030, empower and promote the social, economic and political inclusion of all, irrespective of age, sex, disability, race, ethnicity, origin, religion or economic or other status

We mainly used these targets as our main incentive to approach this problem to analyze how the income disparity between the various percentiles of the population is present and how various factors like education level can impact the financial earning capabilities of an individual.

Apart from these, we also looked at a couple of studies done by the US Census bureau[2] and Pew Research Center[3] which showed there was a clear disparity among the median income between races, which could also be a major factor in contributing to income inequality. So, we also wanted to do an in-depth analysis of how being a person of a different race might attribute to the income level in a particular region of the country.

3. Data

Our primary source of data was from the US Census Bureau, American Consumer Survey. It contained block level as well as place level data of number and type of population for a given geo ID along with numerous income indicators for the neighbourhood such as median income, median rent, mortgaged housing units and many more.

Our data has about 2.2M records and was about 9GB in total size and we used different parts of it for various aspects of our project.

4. Methods

In this section, we first define our workflow followed by a detailed description of approach behind similarity search and different hypotheses.

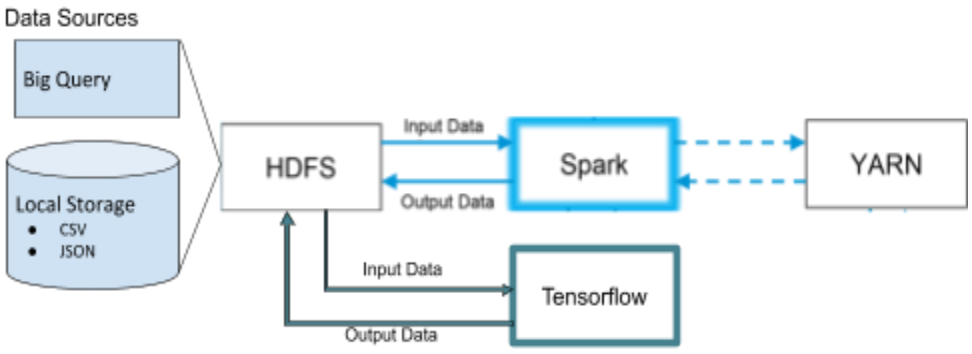


Fig 3.1 Workflow

4.1 Problem definition

Given relational data from the U.S. census bureau we formulated 5 hypotheses based on various social, economic and geographic indicators, followed by multivariate linear regression, hyperplane based LSH and k-means to determine attributes having significant impact on the existing inequality.

4.2 Hypothesis

All hypotheses of this project are centered around income of a geographic block in the United States represented with its unique geoId. Being median income on a geographic level, we can safely assume that this attribute is a part of normal distribution by central limit theorem and weak law of large numbers. Based on this assumption, we determine the income tile of a geoId based on their position in discovered normal distribution and use this for further analysis using t-test. Each and every hypothesis is implemented using apache spark as it enables us to perform our hypothesis on a distributed scale due to the large scale of data. We have kept *alpha(α) = 0.05* and *degree of freedom (dof)* = *n* - 1 - *d*, where *n* is # rows and *d* being # attributes.

One sample t-test = $\frac{\overline{x} - \mu_0}{SD/\sqrt{n}}$ *where* $\overline{x}$ → *sample mean,* μ_0 → *estimated mean of total population, SD → sample standard deviation, n →sample set size*

Paired two sample t-test = $\frac{\overline{D}}{SD/\sqrt{n}}$ *where D (difference distribution) → X-Y,* $\overline{D}$ *→ mean(D), X → sample set, Y → sample set of total population, SD → Standard deviation of D, n → sample set size*

Hypothesis 1: Average growth rate of income per capita of bottom 40tile = National average

For each year, the average income per capita of the bottom 40tile and the total population was calculated. Based on the averages, the year-over-year growth rate was calculated, which results in having two distributions. T-value was calculated by running a one-tailed paired T-test over the two distributions. Consider difference distribution as $(Gr_{40} - Gr_{Total})$ where Gr_{40} → Average growth rate of bottom 40tile population, Gr_{Total} → Average growth rate of total population.

Hypothesis 2: Growth rate of income per capita of bottom 40tile = National average

For each year, growth rates of individual geoids of the bottom 40tile population was calculated. The individual mean of the distributions corresponding to was taken as a valid estimator of the true growth rate of income per capita of the population. Using the estimators of each year from the total population, multiple one-tailed t-tests were run in parallel with the null hypothesis that the growth rate of the bottom 40tile is equal to that of the total population (estimated value).

Hypothesis 3: Tracking growth rate incline of bottom 40tile based in income per capita of 2006

We believe that for the bottom 40tile population, their growth rate should be constant if not growing to be a good indicator of reduced inequalities. So, we picked up the bottom 40tile of 2006 and tracked their income per capita growth rate. We ran multiple one tailed paired T-tests in parallel with the null hypothesis as growth rate of current year = growth rate of next year. Consider difference distribution as $(Gr_{curr} - Gr_{next})$ where Gr_{curr} → Growth rate of current year Gr_{next} → Growth rate of next year.

Hypothesis 4: Growth rate of mean income per capita of bottom 40 tile = top 20 tile
One way of reducing inequality is when the income growth rate of the bottom 40 > top20. This hypothesis is conducted to determine growing inequality between 2 groups using paired two tailed t tests.

Hypothesis 5: Growth rate of educated population with at least Bachelor's degree of bottom 40 tile = top 20 tile. This hypothesis is conducted to determine correlation between income per capita county in respect to its educated population to enable us to have an insight on the extent of impact education has. Only the population with educational qualifications of bachelor or higher are considered.

4.3 Multivariate Linear Regression

To find the essential contributing factor amongst various social, economic and educational indicators, who has positive correlation with median income, we have utilized multivariate linear regression. The following has been implemented using tensorflow due to its high performance on machine learning algorithms.
$y = \beta X$ & $\beta = (X^T X)^{-1} . (X^T . y)$, where y = final output, β represents standardized regression coefficient and X represents input data.

4.4 Finding nth percentile in Hadoop

With data this huge finding elements in a sorted fashion has always been a bottleneck. Since data being a part of normal distribution we propose a unique approach to find the nth percentile of distribution.

1. Find the minimum and maximum of the search domain (analogous to minimum and maximum values of the array in binary search) and the dividing space to be divided into n^{th} tile (analogous to size of array in binary search).
   In our case, the search domain is the income per capita and the dividing space is population.
2. Propose mid-point (or some other ratio) between minimum and maximum as the n^{th} tile.
3. Find if we have underestimated or overestimated the value by running an existing distributed counting algorithm in the dividing space.
   In our case, we count the population below the proposed value.
4. If we have underestimated, then update the minimum to (proposed value + 1). Else, update the maximum to (proposed value - 1).
5. Go to step 2, if max>=min
6. Return the last proposed value as the n^{th} percentile.

It requires log_2*(original maximum - original minimum)* iterations to find the n^{th} tile. In our case, it took 21 iterations to find the bottom 40^{th} percentile

4.5 LSH

Next part of the project involved doing LSH, which we did using Apache Spark and MLLib. After we had pre-processed our data, we concatenated all the features into one column and created hashes using that column. For doing this, we used all the demography related features present in our data to represent the census blocks as hashes. To create these hashes we used a Euclidean distance based LSH

technique called Bucketed Random Projection hash[4]. To generate these hashes, BRP LSH technique used this formula:

$$\left[\frac{h.\,x}{n}\right]$$

Where,
x⟶ feature vectors
v⟶ random unit vector
r ⟶ user-defined bucket length

_c0	features	hashes
490490102102	[28.0,301.0,1993....	[[814.0], [5234.0...
120111103032	[55.0,427.0,1995....	[[1142.0], [7406....
550170111003	[103.0,314.0,1978...	[[-277.0], [2822....
170978649031	[125.0,622.0,1966...	[[885.0], [11512....
360593015003	[48.0,278.0,1943....	[[6757.0], [21529...
371790210071	[56.0,746.0,2002....	[[91.0], [10845.0...
510594602002	[55.0,578.0,1976....	[[2999.0], [15216...

Creating hashes of the features using BRP LSH

The bucket length was used to control the average size of hash buckets (and thus the number of buckets). A larger bucket length (i.e., fewer buckets) increased the probability of features being hashed to the same bucket (increasing the numbers of true and false positives).
However, on further analysis we found out that our data contained overlapping clusters and a nearest neighbor based clustering approach such as LSH would not be viable. So, we used KMeans algorithm.

4.6 KMeans

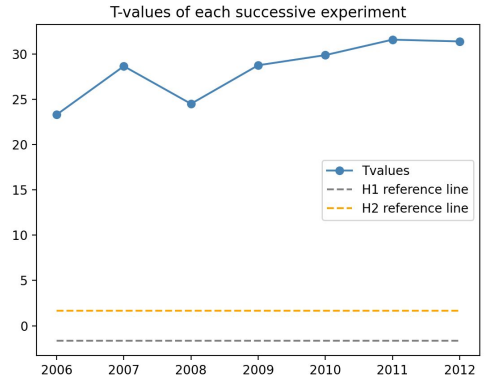
We aimed at analysing inequality within the bottom 40% of the population. For this problem, we wanted to cluster demographically similar blocks that had a median income lower than national median income. In order to cluster these blocks, we represented them in terms of their demographic features such as percentage of people belonging to every race, male/female population etc. Since we had a large number of features, we fit a PCA model to extract important features and ran KMeans on the basis of 15 principal components. However, the clusters formed using this approach did not reveal interesting demographic differences. So, we resorted to clustering on the basis of 7 racial features - number of black, white, hispanic, asian, american indian, 2 or more races, other race people. This methodology gave us clusters that had blocks having similar proportions of different races. For example, cluster 0 contained all the blocks having 73% white, 10% hispanic, 10% black people. To decide the number of clusters, we tried out a different number of clusters and measured “inertia” to find the elbow point on the graph shown below. We then performed analysis of growth of median income across 4 clusters.

5. Evaluation/Results

5.1 Hypothesis

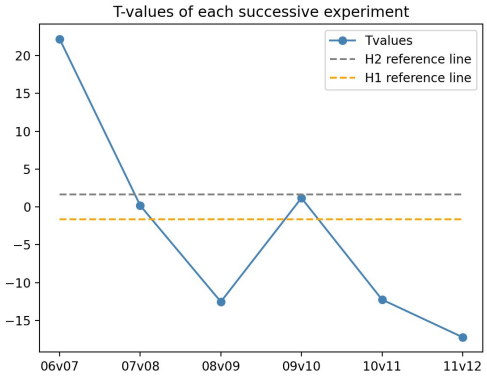
Null Hypothesis	*p-val*	*t-val*	*Retain/ Reject*
Average growth rate of income per capita of bottom 40 tile = National average	0.41	0.23	Retain
Growth rate of mean income per capita of bottom 40 tile = top 20 tile	0.89	-1.4	Retain
Growth rate of educated population with at least Bachelor’s degree of bottom 40 tile = top 20 tile	0.44	0.14	Retain

Hypothesis 2 Result



H2: GR of bottom 40 tile > av GR of total pop
As T-value has been consistently high, we can conclude that the income per capita growth rate of the bottom 40tile is more than that of the total population.

Hypothesis 3 Result



H1: GR of curr year < GR of next year
H2: GR of curr year > GR of next year
We see that even though the growth rate of income per capita was increasing, after 2007, it started decreasing.

5.4 Multivariate Linear Regression

We analyzed data from 2010 to 2018 and our findings indicate that out of numerous social, economic and education related attributes, counties with high numbers of people with educational qualification of master's degrees or higher have a positive correlation with national median income. This indicates the importance of education in bridging the income inequality gap between bottom 40 and national average.

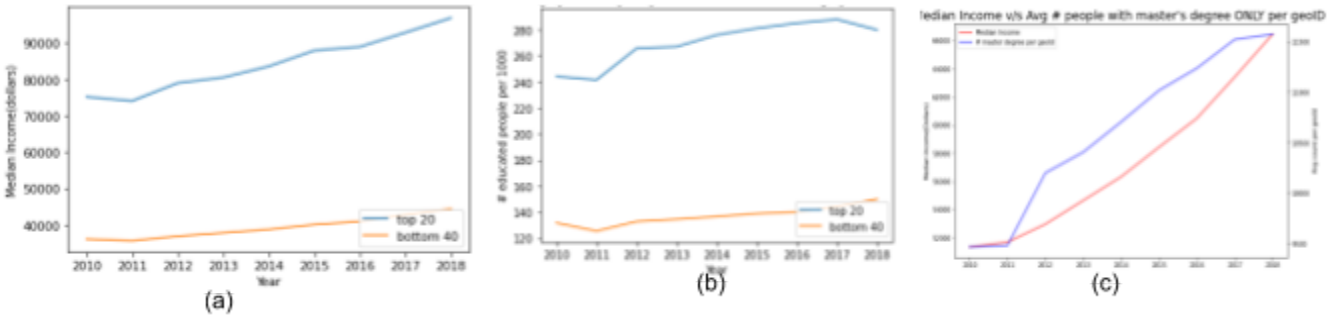


Fig 5.3: (a) YOY avg income of bottom 40 vs top 20, (b) YOY avg educated population of bottom 40 vs top 20 and (c) correlation between population with master's degree or higher and national median income

5.5 Similarity Search and Cluster Analysis

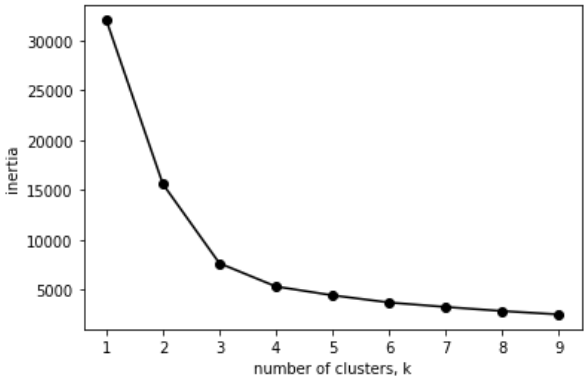
After we had formed a pipeline to compute the hashes for our features, we wanted to test the optimality of the hashes. Since we had to define an L2 distance metric to keep as a threshold for selecting viable buckets, we could test the working of our BRP LSH model.

For the similarity search module, we had to do an approximate self join on our data so that the similarities between the hashes could be computed. The values were selected based on an L2 distance threshold to be identified as a similarity pair. After doing some experiments, we found out that reducing the dimensionality of our data helped improve performance of the similarity search model a lot.

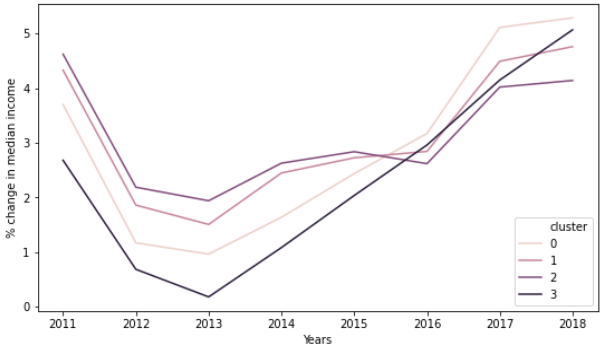
datasetA	datasetB	distCol
[421279603001, 10...	[360099611004, 10...	75.07329751649384
[201079552002, 11...	[261590101005, 11...	93.03762679690406
[380779714002, 79...	[420391119001, 81...	58.532042506647585
[460619641003, 11...	[180950101001, 11...	85.17628778010932
[295101083002, 89...	[245102007015, 87...	94.2178327069775
[551239607004, 11...	[421150326001, 11...	79.0
[360010127003, 58...	[172010015002, 56...	98.2496819333274

Computing Similarity after reducing dimensionality in data

We also computed the n-nearest neighbors by comparing the hashes across our data with a key value which we wanted to do to form clusters of our data. But due to overlapping clusters, this was not feasible so we did cluster analysis after making use of KMeans approach.



Measuring "inertia" for different numbers of clusters formed using KMeans. As we can see, the curve flattens after k=4 indicating no significant improvement.



% growth rate of median income of different clusters across different years. The growth rate has become less unequal in recent years.

6. Conclusion

Income per capita growth rate of the bottom 40tile has been consistently high as compared to the national average. However, when we dig down deeper, we see that the increase in growth rate has slowed down over the years and even decreased. We see that higher education has a positive impact on income. Policies promoting education, specifically Master's degrees, should reduce inequality among the masses. Also, we found out that while LSH was very useful in analyzing our distributed data, it was not viable for the formation of clusters which we wanted to do to analyze similar trends and growth rates. So we used the tried and tested KMeans method and found that there are positive signs as the growth rate has started to become less and less unequal among the various clusters of people in recent years.

7. References

[1] Classifying countries by income, Espen Beer Prydz, Divyanshi Wadhwa, WorldBank Data Topics
[2] Income and Poverty in the United States, US Census, 2017, p-5
[3] https://www.pewresearch.org/fact-tank/2020/02/07/6-facts-about-economic-inequality-in-the-u-s/
[4] https://spark.apache.org/docs/latest/ml-features.html#lsh-operations
[5] Detecting Abuse at Scale:Locality Sensitive Hashing at Uber Engineering, Engineering Blog, DataBricks
[6] American Consumer Survey (ACS), United States Census Bureau