Exploring the disparity in basic amenities on the basis of caste in the Indian community.

Authors Name/s Syed Abid Abdullah
Department of Computer Science
BITS Pilani Hyderabad Campus
Hyderabad 500078, India
Email id: f20150562@hyderabad.bits-pilani.ac.in

Authors Name/s Tej Sukhatme
Department of Computer Science
BITS Pilani Hyderabad Campus
Hyderabad 500078, India
Email id: f20180020@hyderabad.bits-pilani.ac.in

Authors Name/s Akash Srivastava
Department of Computer Science
BITS Pilani Hyderabad Campus
Hyderabad 500078, India
Email id: f20180241@hyderabad.bits-pilani.ac.in

***Abstract*—A common social outlook in the Indian community is that the caste division produces disparity in basic amenities. Although the divide is historically rooted, the common opinion is put through research and surveys. This electronic document entails various methodologies and procedures adopted in and on the path to inferring information regarding the topic.**

Keywords*—*Data cleaning or Data cleansing; Attribute subset selection; Data reduction

I. Introduction

In the context of the Indian society, caste is defined to be a Hindu hereditary class of socially equal persons, united in religion and usually following similar occupations, distinguished from other castes in the hierarchy by its relative degree of spiritual purity or pollution. [1]

In modern India, the Indian government introduced a categorization scheme in which the untouchable castes were categorized as scheduled castes (SC), the backward tribes were categorized as scheduled tribes (ST) and the disadvantaged castes as other backward castes (OBC). [2] This study could prove as a rather interesting avenue against the problem of disparity due to caste.

II. Problem Definition

The proposition under inspection is that caste still affects the general lifestyle and facilities enjoyed by an objective person in India.

III. Data Preprocessing

A. Data Cleaning

Data cleansing or **data cleaning** is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record, set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores.

The dataset originally had several human errors that needed correction. Things like there being extra commas in the middle of district names. which led to new rows being formed in the .csv file. These errors were tackled using the bash script and regular expressions. Luckily there were no missing values so none of the rows were deleted.

B. Attribute Subset Selection

The data set may have a large number of attributes. But some of those attributes can be irrelevant or redundant. The goal of attribute subset selection is to find a minimum set of attributes such that the dropping of those irrelevant attributes does not much affect the utility of data and the cost of data analysis could be reduced. Mining on a reduced data set also makes the discovered pattern easier to understand.

Domain knowledge was utilized to remove features that were irrelevant, for example attributes like: status, SC concentrated, ST concentrated, etc

C. Data Reduction

Since 16,00,000+ rows were originally present we combined all the villages into their corresponding districts and added all the population values to end up with a database with around 600 entries.

IV. Visualization

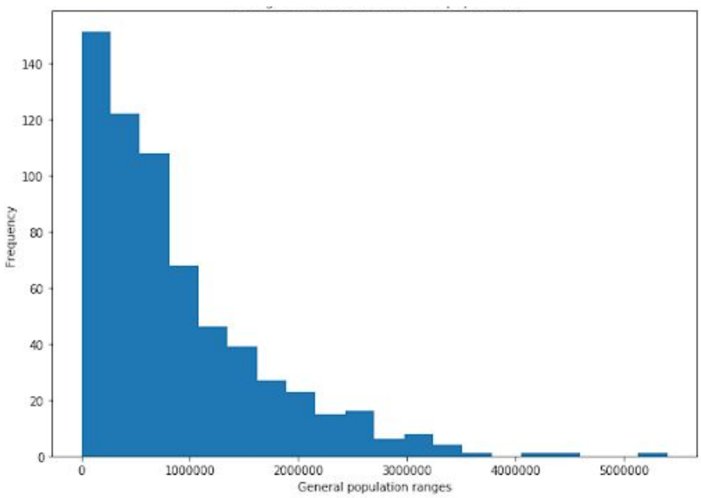

Fig 1: Histogram of General Covered Population

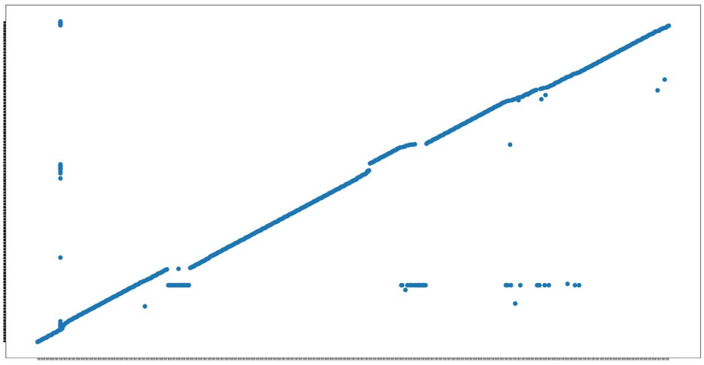

Fig 2: SC vs ST correlation data

References

[1] Susan Bayly.Caste, society, and politics in India from the eighteenth century to the modern age, volume 3. Cambridge University Press, 2001.

[2] Sheth, D. L. (1987). Reservations policy revisited. *Econ. Polit. Wkly.* 22, 461957–461962.