THE ACHIEVERS

Siddharth S Desai, Rushikesh Jadhav, Maj. Abhishek Bhadouriya

➢ INTRODUCTION:

Motivation: MNREGA is one of the largest public employment programs in the world, and its data can provide insights into the socio-economic conditions of rural people in India. It can be the potential dataset to study the policy making and the development efforts made my Indian Gov., which will directly impact people's lives.

Problem Statement: How MNREGA helps for Below Poverty Line households (Scheduled Caste)

➢ DESCRIPTION OF DATASET:

The Dataset used is **MNREGA** (Mahatma Gandhi National Rural Employment Guarantee Act, 2005) Dataset. It Contains Information about the Scheme by the Government of India which guarantees seasonal / temporary employment at once a year in rural areas. The data is provided for a ten-year period of 2011-2020.
Dimensions of Dataset and Null Value Count: 6858 × 47, 53.

➢ DATA PREPROCESSING:

Data Preprocessing is the Stage where almost 70% of the time is required. It includes cleaning of dataset for further analysis.

Dropping Irrelevant Columns: It's better to drop few columns which are of no use. We dropped 3 columns:

- **Country**: Dropped this column because the variance of this column was zero, i.e., all the rows contained only one value India. So, it would have used undue space.
- **Rowid**: Dropped RowId because as we change our csv to Data frame, it already comes with the Rowid starting from 0, So this column becomes redundant and hence we dropped it.
- **Year**: This Row is of type Object with the Textual data. Example entry, "Financial Year (Apr - Mar), 2011", which was of no use because we had another column "Yearcode" with the entries in int64 format which had the particular year, so Year Column becomes redundant and hence dropped it.

Handling Null Values: Null Values are present in in 6 columns.

- **Method 1: Dropping Null Values:**
  This strategy can be effective in removing incomplete observations, it can lead to data loss.
  In MNREGA, the rows which are having Null values are coming from few Union Territories if we drop the rows, then we will lose rest of the Information of that Union Territory, so Mean/Median/Zero Imputation will be better in this case.
- **Method 2: Zero Imputation:**
  Zero imputation replaces null values with the zeros directly.
  When we closely looked at the data, we found out that the Row which has NANs, all the columns of that row have zeros, so imputing it with Zeros would make sense here.
  We have imputed the NANs with zeros to preserve the dimensions of the dataset, Here even dropping the NULL values rows would have helped because Null values are present in those rows whose all entries are zeros, but we would have lost data for that State and the mean/median is zero for that column.

Checking For Outliers:

- **Using Box Plots [2] [Fig1]:** Box plot provide a visual summary of the distribution of data. We can see points outside the whiskers of the box plot which means data is distributed away from Median. The data is Positively skewed so we got this kind of box plot.
- **Using Scatter Plots [3] [Fig2]**: Even Scatter plots make a clear visualization that there are Outliers in the Dataset., Some of the points in Dataset are far away representing Outliers.

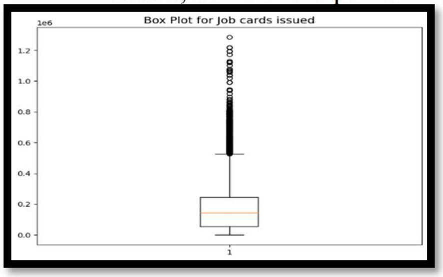


Fig1: Box Plot of single column

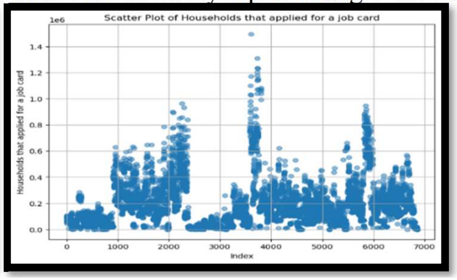


Fig2: Scatter plot of a single Column

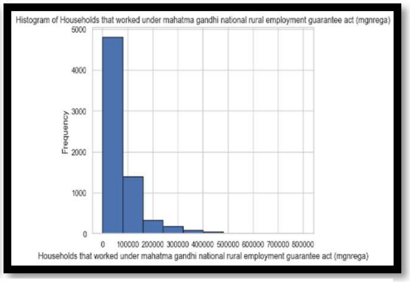


Fig3: Distribution of Data

Handling Outliers: The dataset is of MNREGA, so it has variance in each column and dropping those outliers will result in losing out Important Information about the columns. So, we will be working with those outliers, expecting that we can find some information out of those outliers.

Distribution of Dataset [Fig3]: Our data is Positive-skewed, i.e., most of the values are near to zero.
Scaling And Standardization: Some of the columns in the dataset are having large range of values, and some columns have less range, so the model needs a scaled data, which means that all the columns will have the range in between 0-1.

- **Standard Scaler:** We used the inbuilt Standard Scaler$\left[z = \frac{x-\mu}{\sigma}\right]$ module to scale the data, which just brings values between 0-1. Our data is Exponentially decaying and with outliers, so this scaling method didn't work when doing PCA and Clustering it just brought data between 0-1, the data was not normally distributed yet.
- **Min-Max Scaler**: We then went with Min-Max Scaling, it also performs bad with Exponentially decaying data and data which doesn't follow Gaussian Distribution.

Note: We have scaled the data wherever necessary. The scaling is not done explicitly in the preprocessing steps, because we need to visualize the data with the actual values and not the standardized data. We have standardized it before applying Clustering and Dim - Reduction.

➢ DESCRIPTIVE ANALYSIS:

- **Job Cards Issued Over the Years [Fig4]:** The plot shows that there is a sudden increase in the number of Job cards issued in 2020, which was the start of COVID-19, resulting in a greater number of people getting jobless, and applying for MNREGA Schemes.
- **Distribution of Household That applied for Job Card by State [Fig5]:** West Bengal has the highest number of applicants, followed by Bihar. These states are predominantly rural with heavy agriculture dependence on the state economy. These are the least Urbanised States in India. So, the plot clearly states that UP and Bihar people need more job opportunities as compared to other states on India.
- **Job Cards Applied vs Issued [Fig6]:** We can see in the Figure that almost everyone who applied for job card are issued a Job card, which says that the Government of India is successfully in creating jobs for the needy, and the scheme is really a success in Indian market.
- **Days worked by Women [Fig7]:** All the years before 2019, the rate of growth of working women was less. The sudden spike was seen in 2020. This was the starting period of Covid-19, which gives us the information that there was a pressure on the women also to go to job as the covid-19 impacted most of the households. Women are also participating in the MNREGA schemes.

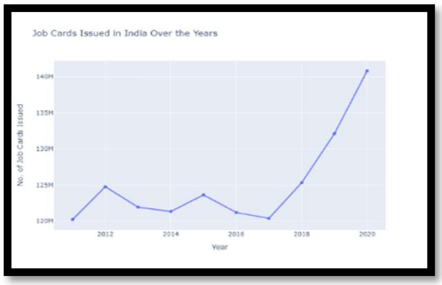


Fig4: Job cards issued over the years.
.

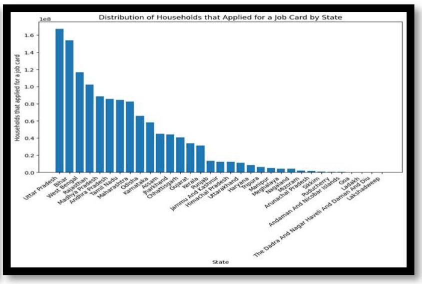


Fig5: People that applied for Job.

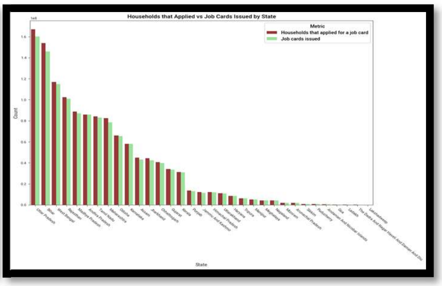


Fig6: Job Cards Applied [Brown] vs issued [Green]

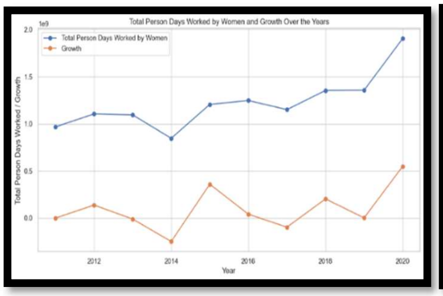


Fig7: Days worked by women.

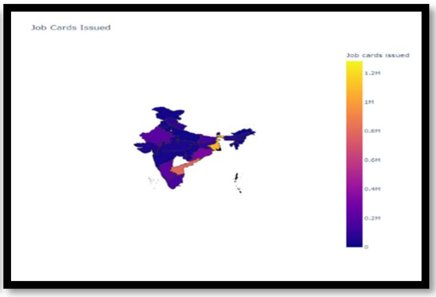


Fig8: Job card issued in each state

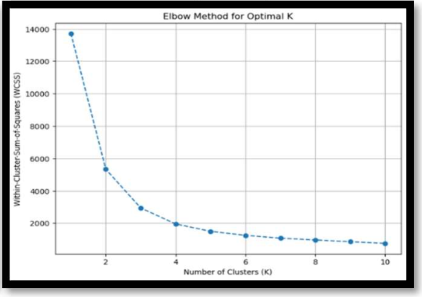


Fig9: Elbow Method

- **CLUSTERING ANALYSIS:** Using Elbow Method **[Fig9]**, we can see that 3 clusters can be formed out of the data-points.

K-means [Fig10]: Applying K-means didn't give satisfactorily results because the shapes of the clusters are not globular. K-means assumes that clusters are globular. So, K-Means didn't work here.

Spectral Clustering [Fig11]: Gave somewhat better results than K-means but didn't handled outliers well. Spectral clustering needs Fully connected graph. The data here may not be full connected and hence spectral embedding didn't work as expected. Affinity matrix used was Nearest_neighbors.

Spectral Clustering with rbf [Fig12]: Trying Change of affinity from nearest_neighbors to "rbf" also didn't work, it improved a bit. It was able to identify the outliers and make a separate cluster for it.

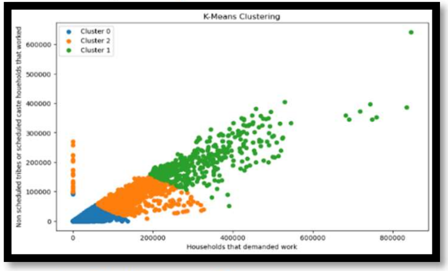


Fig10: Kmeans Clustering

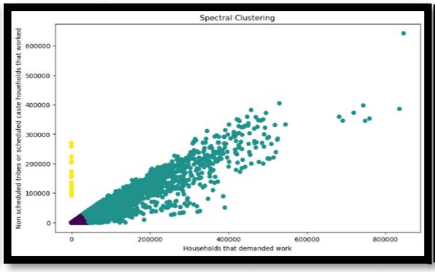


Fig11: Spectral Clustering with affinity=nearest_neighbors

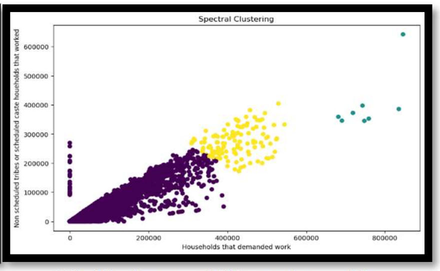


Fig12: Spectral Clustering with affinity=rbf

Agglomerative Clustering [Fig13]: This clustering doesn't assume inherent shape of clusters, so we tried applying this. But here agglomerative clustering was not able to cluster with default parameters as the results vary based on linkage criteria.

Agglomerative Clustering with parameter change [Fig14]: We tried different parameters, but the parameters which worked for out data were **single linkage** with affinity matrix as **'Euclidean'**. Choosing Manhattan as affinity also gave the same results. Agglomerative algo didn't worked but change in affinity gave better results.

DBscan [Fig15]: Normal DBscan worked as Kmeans but tweaking the values of **eps** and **MinPts** resulted in a cluster with a **silhouette score of 0.85**. Density based clustering worked well here because this algo works for any arbitrary shape. It also works well with Noise and Outliers. It cluster's Outliers separately.

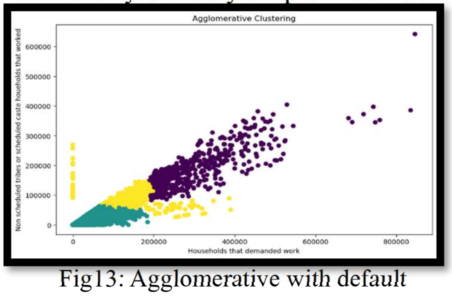


Fig13: Agglomerative with default values

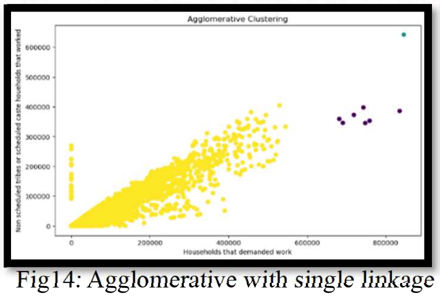


Fig14: Agglomerative with single linkage and Euclidean affinity

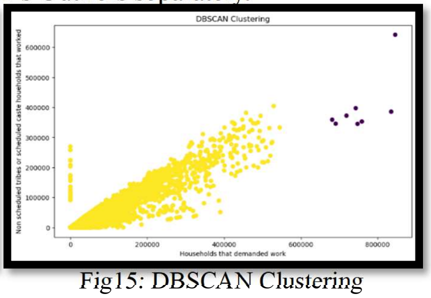


Fig15: DBSCAN Clustering

Algorithm	K-means	**Spectral Clustering** with nearest neighbour	**Spectral Clustering with rbf**	**Agglomerative Clustering** with default parameters	**Agglomerative Clustering with** single linkage	DBscan
Silhouette score	**0.744**	**0.256**	**0.771**	**0.546**	**0.844**	**0.85**

CLUSTERING ON TSNE DATA: Fig21 shows the clusters formed using K-Means with K= 4.

Insights:

1) Cluster coloured in **purple** [Top] denote that these are the states where the MNREGA participation is least. This cluster is mostly of Union Territories like Lakshadweep, Andaman and Nicobar etc, where the awareness of scheme is not done nicely. Government needs to do some awareness camps in these Union territories so that more participation can come out of these Union-Territories. The ratio of job card applied vs issued is less in these parts of India [Fig6]
2) Cluster coloured in **orange** are the states where Mnrega is successful but is not able to target much population in those areas. The participation is less as compared to overall population. Government needs to bring in more and more people under the scheme so that more households can take advantage of scheme.

3) Cluster coloured in **Blue** are the states where this scheme is highly successful with maximum participation of the population in the schemes and maximum people are taking advantage of the scheme proposed.
4) Cluster coloured in **Yellow** are the Eastern States which are less populated, but the scheme was successful in creating jobs. The participation of the women was seen more in these states.

➢ DIMENSIONALITY REDUCTION:

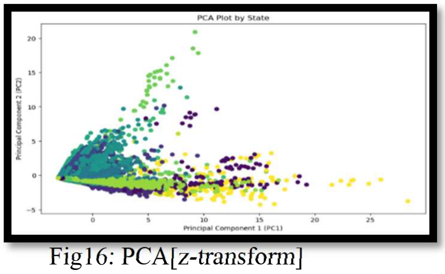
PCA (z-transformation) [Fig16]: We standardized the data using z transform $[z = (x - \mu)/\sigma]$. The features in original data were not able to form clusters as the data was Exponentially decaying. After reducing the dimensions of data to 2, there was not any significant change in the plots. We kept 3 PC but even that didn't work because of complex distribution and z transform didn't help to make the data Linearly Separable.

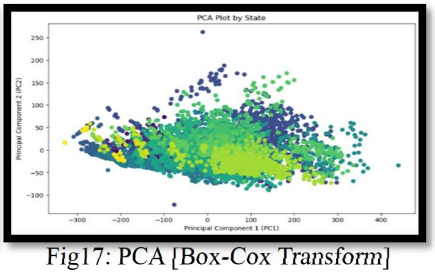
PCA (Min-Max scaling): This worked like z-transform, because min-max scaling is not efficient when data is like exponential decay.

PCA (Box-Cox Transformation) [Fig17]: Applying box-cox to the data changed the data almost near to normal distribution, but it was able to do it for certain columns. Applying PCA on this transformed data, gave plots, but we couldn't see any clusters in the dataset. Box-Cox is effective when the data follows a close-to-normal distribution. But our data has a complex distribution [Exponentially decaying] so box-cox failed.

PCA (Log Transformation) [Fig18]: This is effective technique when the data is having some complex distribution like skewness. Log transformation changed the distribution almost to normal distribution. Applying PCA on it gave the results as in Fig18.

Tried **Multi-dimensional Scaling (MDS)** also but the results were not satisfactorily due to structure of the data(Exponentially decay like) and the presence of outliers in the data.





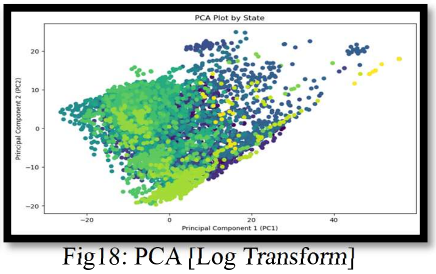
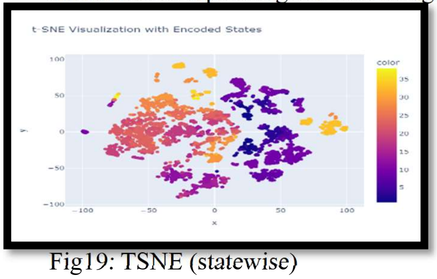
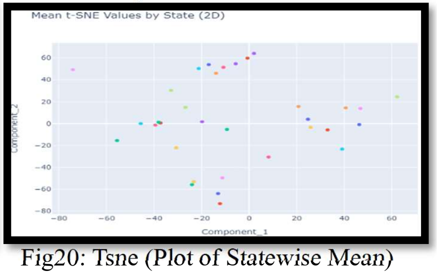


Fig16: PCA[*z-transform]*   Fig17: PCA *[Box-Cox Transform]*   Fig18: PCA *[Log Transform]*

TSNE: Applying TSNE with the log transformed Data gave the result as in [Fig19]. Our data had non-linear and very complex structure, hence TSNE seemed to be effective here. Fig19 is plotted statewise. To visualize it better we grouped data by State and took the statewise-mean. In **Fig20** we can see the data is showing some statistics about itself. We were successful in separating the data using T-SNE. This proves that TSNE works well with complex structures.





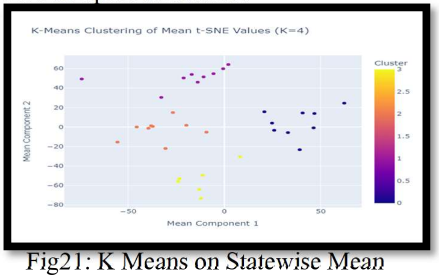


Fig19: TSNE *(statewise)*   Fig20: *Tsne (Plot of Statewise Mean)*   Fig21: *K Means on Statewise Mean*

➢ ISSUES/CHALLENGES FACED:

1. Curse of Dimensionality: Not able to separate data to learn insights, even after applying Dimensionality Reduction Technique (PCA, MDS). No Visualization of data despite of reducing dim using PCA and MDS.
2. Elbow plot needed specific version of threadpoolct, it was giving error with the version installed in the system. We were not able to find out what error was it exactly. Later changing the version helped.

- **Reason why Dim-Reduction didn't work with PCA:**

1) Highly overlapped data points. [Can be seen visually by scatter plot (Fig2)]
2) Highly Linear correlation between the columns. [Plotted Correlation Matrix and values are close to one], and PCA assumes no linear relationship between the columns. So PCA failed.
3) Many outliers: The data contains outliers, which we can't drop as it contains useful info. and insights.
4) Non-Linear data: PCA was not able to separate the data and show some clusters. Hence, we can conclude that our data is Non-Linear.

- **Reason why clustering didn't work with PCA:** We were not able to separate data using PCA because the data was not linearly seperable. Hence, we tried brute force (i.e Scattering between relevant col.) and due to high correlation, we ended with bad cluster formation. With TSNE we were able to make the data Linearly seperable at some extent, which helped us in visualizing and understanding some insights.
- **Method to tackle this problem:**

Feature Scaling and Transformation: We tried using box-cox and log transformation. But it failed due to the structure and complexity of data i.e., Highly overlapped data points and exponentially decaying distribution
Using brute Force or Manual Inspection: It may not be feasible method in real life, but when Dimensionality Reduction didn't help to visualize and cluster formation, then we came up with a method where to visualize the data, we take 2 relevant columns from the data and see the distribution of those columns whether it can give us some insights. Similarly, we did by selecting 3 relevant columns and visualizing it.
Using TSNE: Through above analysis we have concluded that, PCA is popular Dim-Reduction technique, but sometimes its better to try other Dim-Reduction Techniques also.

➢ **MANUAL INSPECTION:** We took relevant columns which are useful in our problem statement. 2D and 3D Scatter plot of each combination of columns gave us an idea where we can extract some data out of it.

Objective: Study about SC households that were allocated work. We chose columns related to our Objective. After applying different clustering algos, the best result was given by DBSCAN shown in **Fig15.**

- The cluster coloured in Yellow gives us the information that out of the total Households that demanded work, maximum number of works was allotted to SC families. It states that this community needs jobs.
- The cluster in purple colour is the cluster of outliers. This cluster tells, even if number of Households that demanded work increase above 6 lakhs, still the jobs given to SC is limited, which directly states that the number of jobs for SC are fixed to some quantity up to 4 lakhs.

➢ **RESULT OF ANALYSIS USEFUL FOR INDIAN GOVERNMENT:**

- SC Households are more in need of the scheme, so govt should focus on creating more job opportunities for them as well as create awareness of the scheme so that more Households can take advantage of it.
- The Union territories need to be focused more as far as this scheme is concerned, in terms of creating awareness amongst the people about the scheme and bringing more participation from these locations.

➢ **CONCLUSION / INSIGHTS GAIN:**

- There are some datasets where even after applying PCA, we need not form any clusters in the dataset. The data cannot be visualized even after applying PCA.
- The box plot sometimes doesn't talk about the outliers. According to **Fig1**, all the points are in only one Quartile, which meant that data is distributed in that quartile only. After plotting scatter plot, we saw the outliers visually. But it gave an idea about the distribution of the data.
- Spectral clustering is very sensitive to affinity. With RBF kernel, spectral clustering isolates outliers, leaving clusters with numerous observations [Fig12].
- PCA works with any data distribution, but if data is normally distributed, then PCA works better given that there are no outliers in the dataset and no linear relationship, between the data.
- In real life data need not be as clean as the Kaggle datasets are, we must try and test different techniques and see which suits our data well.

➢ **SELF-REFLECTION:** The project taught us to find insights from the data, which may be important for further analysis. We were able to enhance our critical thinking, which helped us to reach at the conclusion of applying Manual Inspection of data.

➢ **REFERENCES:**

[1] Kaggle Dataset: Sumedha Poonia."Government Aided Employment in India (MGNREGA) Kaggle. Link

[2] Sweetviz Documentation. "Sweetviz" (Accessed October 25, 2023). Link

[3] Plotly Documentation. "Plotly" (Accessed October 25, 2023). Link

[4] Un-Mapped Carto. "States India Map" (Accessed October 25, 2023). Link

[5] Chat GPT.