॥वसुधैव कुटुम्बकम्॥

SYMBIOSIS

INSTITUTE OF TECHNOLOGY

Machine Learning Report

Submitted To:

Dr. Harshala Shingne

Associate Professor

Department of Computer Science and Engineering Symbiosis

Institute of Technology, Nagpur

Submitted By:

PARIKSHIT ABUJ

Semester: VII

Section: A

PRN: 22070521009

INDEX

Table of Contents

EDA Report: Direct Benefit Transfer (DBT) Dataset

This report includes data cleaning and exploratory data analysis of the district-wise DBT dataset in India. Key analyses include missing value handling, state-wise and district-wise distributions, trends over years, and correlation between transaction volume and amount transferred.

Dataset Overview

The dataset used in this analysis provides a district-wise record of Direct Benefit Transfer (DBT) activities across various Indian states. It includes both monetary transfer values and the number of transactions performed under DBT schemes for each district and financial year. The dataset is sourced from government records and serves as a crucial input for analyzing welfare scheme distribution patterns.

Total Records: 3825 Columns:

- Financial Year (fy)
- State Name, State Code
- District Name, District Code
- Total DBT Transfer (INR)
- Number of DBT Transactions

2.1 Column Name Standardization

Standardizing column names ensures uniformity in naming conventions, which improves readability and simplifies code development. Converting names to lowercase with underscores (`snake_case`) avoids conflicts in programming environments like Python and helps maintain consistency during data manipulation.

2.2 Handling Missing Values

Missing values can distort statistical analysis and visualizations. Removing or imputing such entries is crucial to maintaining data integrity. In this case, records missing crucial transaction and transfer data were removed to avoid bias or misinterpretation in the analysis.

2.3 Removing Duplicates

Duplicate records may arise due to repeated entries during data collection. They can falsely inflate the dataset size and skew results. Therefore, deduplication is a necessary step to ensure each data point is unique and meaningful.

2.4 Data Type Conversion

Data stored in incorrect formats (e.g., numbers as strings) can hinder calculations and analysis. Converting columns into appropriate types (e.g., `int`, `float`, `category`) allows efficient memory usage and enables correct function application during analysis.

2.5 Final Dataset Dimensions

Understanding the final shape of the cleaned dataset (number of rows × columns) confirms the scope of data available for analysis and gives an overview of its structure.

- Standardized column names to lowercase with underscores
- Removed 121 rows with missing values in 'total_dbt_transfer' and 'no_of_dbt_transactions'
- Dropped duplicate records (if any)
- Converted data types for numerical and categorical fields
- Final dataset shape: 3825 rows × 8 columns

NLP

To gain linguistic and regional insights from the *district_name* field, Natural Language Processing (NLP) techniques were applied.

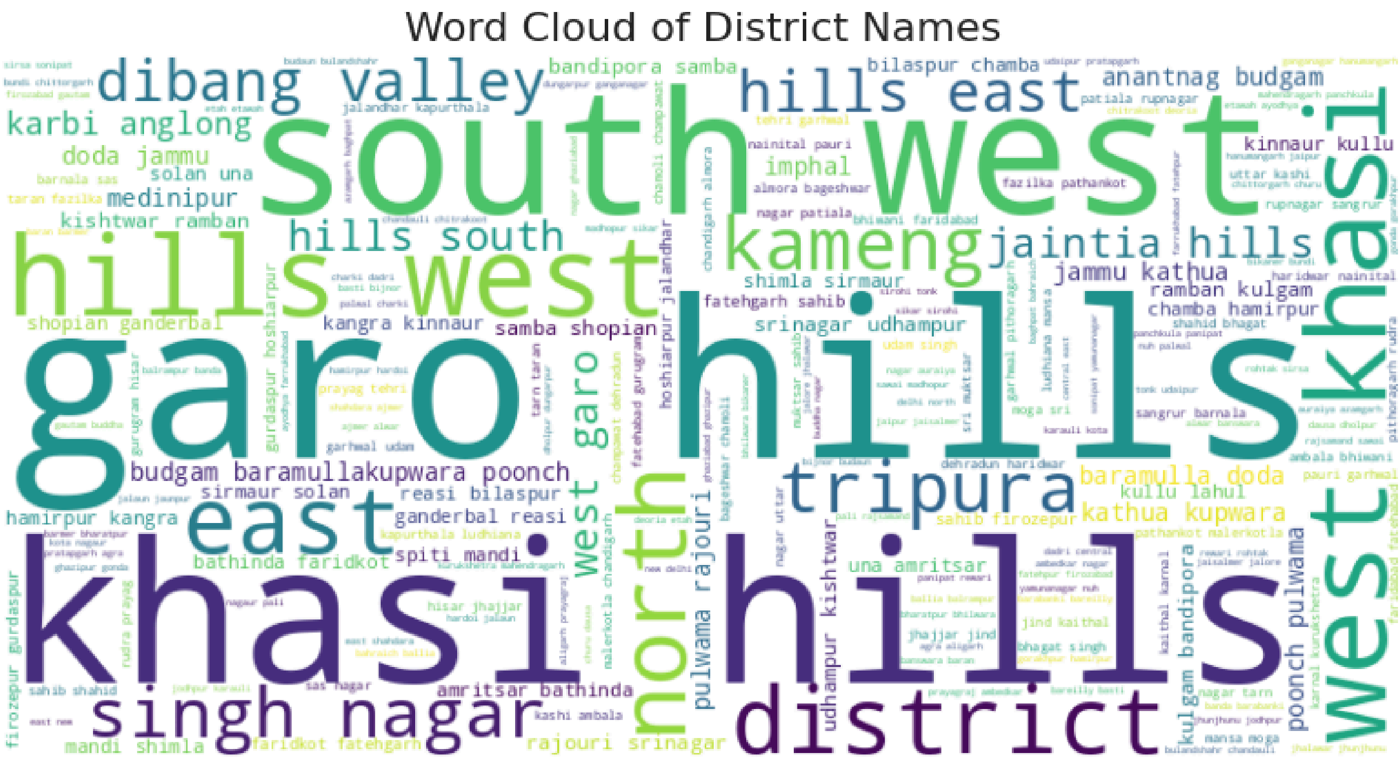
1. District Similarity Analysis:
   Using text embeddings, the system found districts most similar in name to "Aurangabad," revealing several matches across Maharashtra and Bihar. This shows that multiple districts across different states share identical or phonetically similar names, which can influence text-based clustering or data grouping.

2. K-Means Clustering on District Names:
   Clustering based on textual similarity grouped districts with similar name structures (e.g., those ending with "pur", "garh", or "nagar"). This helped identify common naming conventions across states. A sample output shows that districts within Jammu & Kashmir often cluster together due to shared suffixes and linguistic roots.

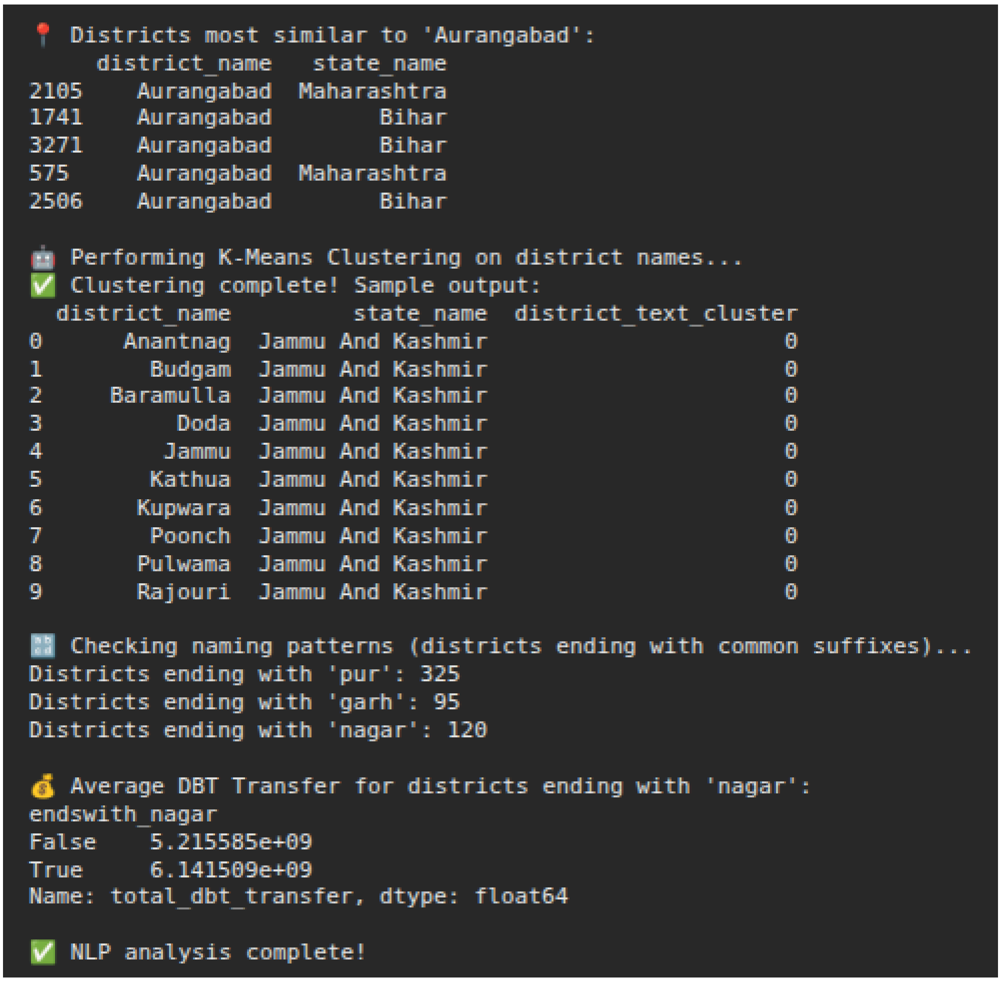
3. Suffix Pattern Analysis:
   The analysis detected 325 districts ending with *"pur"*, 95 with *"garh"*, and 120 with *"nagar"*. These suffixes reflect cultural and linguistic patterns in Indian district naming. Further, it was found that districts ending with "nagar" have a higher average DBT (Direct Benefit Transfer) amount compared to others.

4. Word Cloud Visualization:
   The Word Cloud of District Names (shown below) highlights the most frequent words used in district names. Words like *"hills"*, *"north"*, *"south"*, *"garh"*, and *"nagar"* dominate, indicating strong regional naming trends.

5. NLP Results Snapshot:
   The figure below summarizes the output of the clustering, pattern detection, and DBT transfer comparison.



NLP ON DATASET



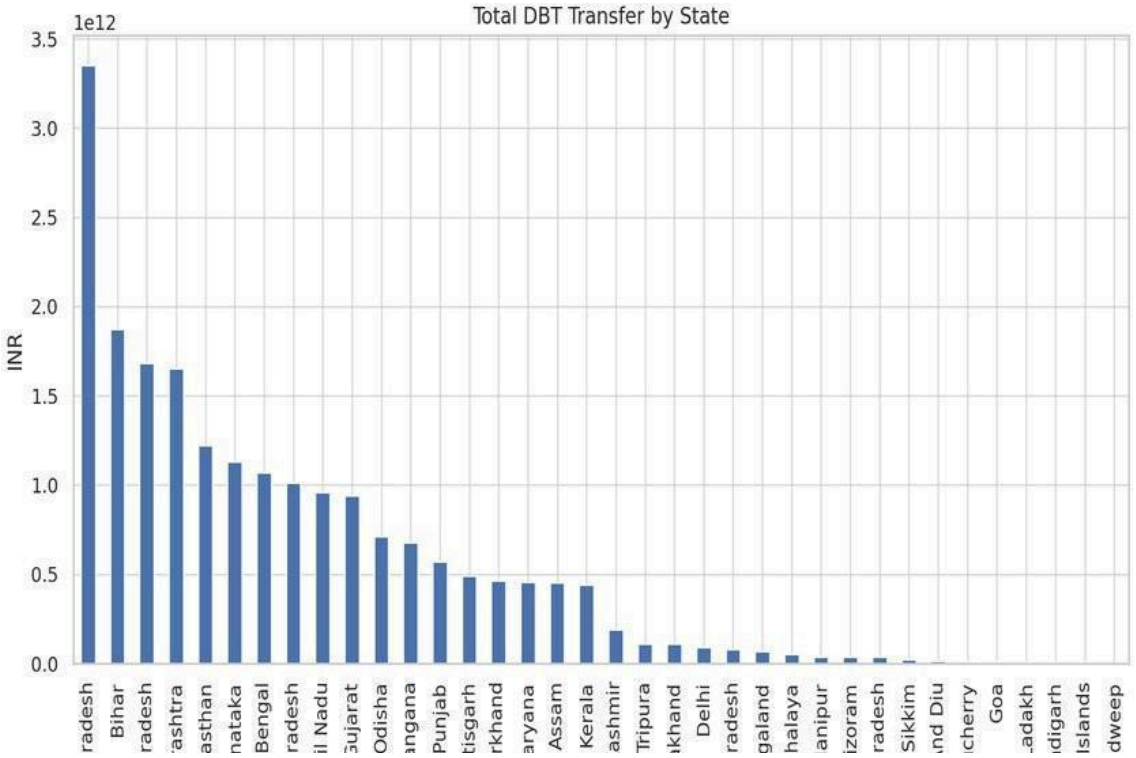
```
📍 Districts most similar to 'Aurangabad':
     district_name   state_name
2105    Aurangabad  Maharashtra
1741    Aurangabad        Bihar
3271    Aurangabad        Bihar
575     Aurangabad  Maharashtra
2506    Aurangabad        Bihar

🤖 Performing K-Means Clustering on district names...
✅ Clustering complete! Sample output:
  district_name         state_name  district_text_cluster
0      Anantnag  Jammu And Kashmir                      0
1        Budgam  Jammu And Kashmir                      0
2     Baramulla  Jammu And Kashmir                      0
3          Doda  Jammu And Kashmir                      0
4         Jammu  Jammu And Kashmir                      0
5        Kathua  Jammu And Kashmir                      0
6       Kupwara  Jammu And Kashmir                      0
7        Poonch  Jammu And Kashmir                      0
8       Pulwama  Jammu And Kashmir                      0
9       Rajouri  Jammu And Kashmir                      0

🔡 Checking naming patterns (districts ending with common suffixes)...
Districts ending with 'pur': 325
Districts ending with 'garh': 95
Districts ending with 'nagar': 120

💰 Average DBT Transfer for districts ending with 'nagar':
endswith_nagar
False    5.215585e+09
True     6.141509e+09
Name: total_dbt_transfer, dtype: float64

✅ NLP analysis complete!
```

CONCLUSION ON NLP

The NLP-based district name analysis revealed strong linguistic and regional patterns, helping correlate naming structures with DBT distribution trends and providing insights useful for regional policy analysis.

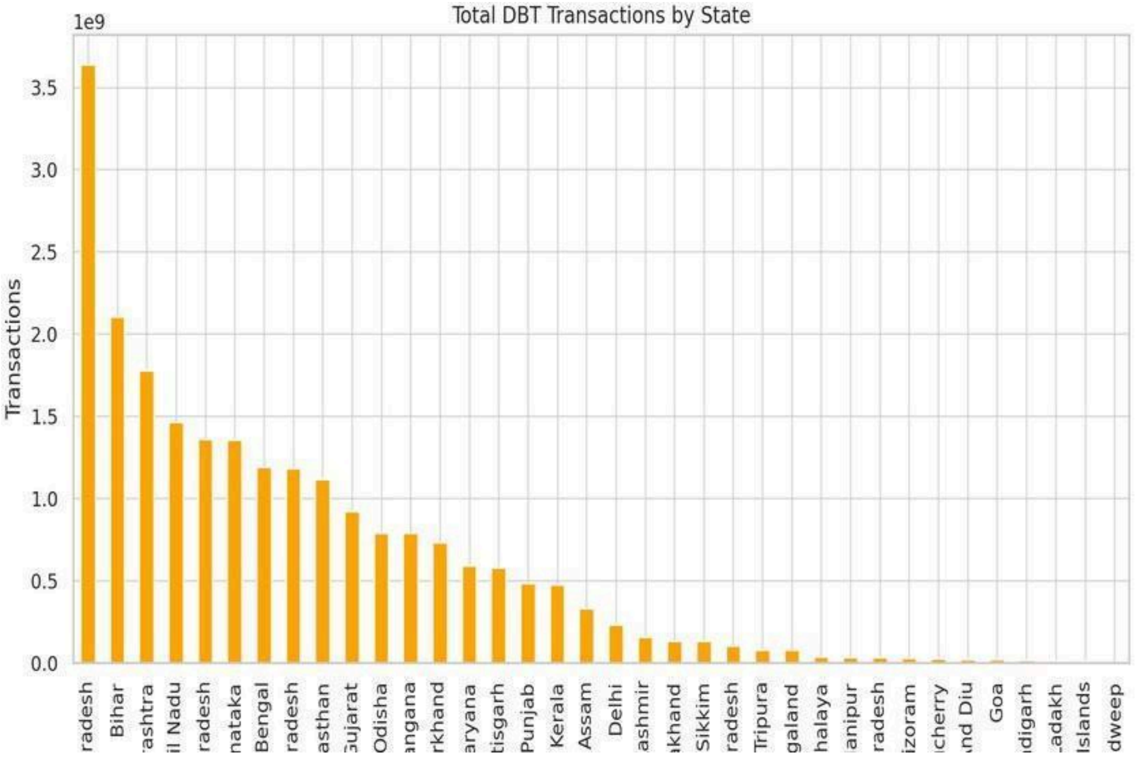
Total DBT Transfer by State

This analysis shows how much money has been transferred under DBT schemes across each state. It helps identify regions receiving the highest government support, enabling comparison and policy evaluation.



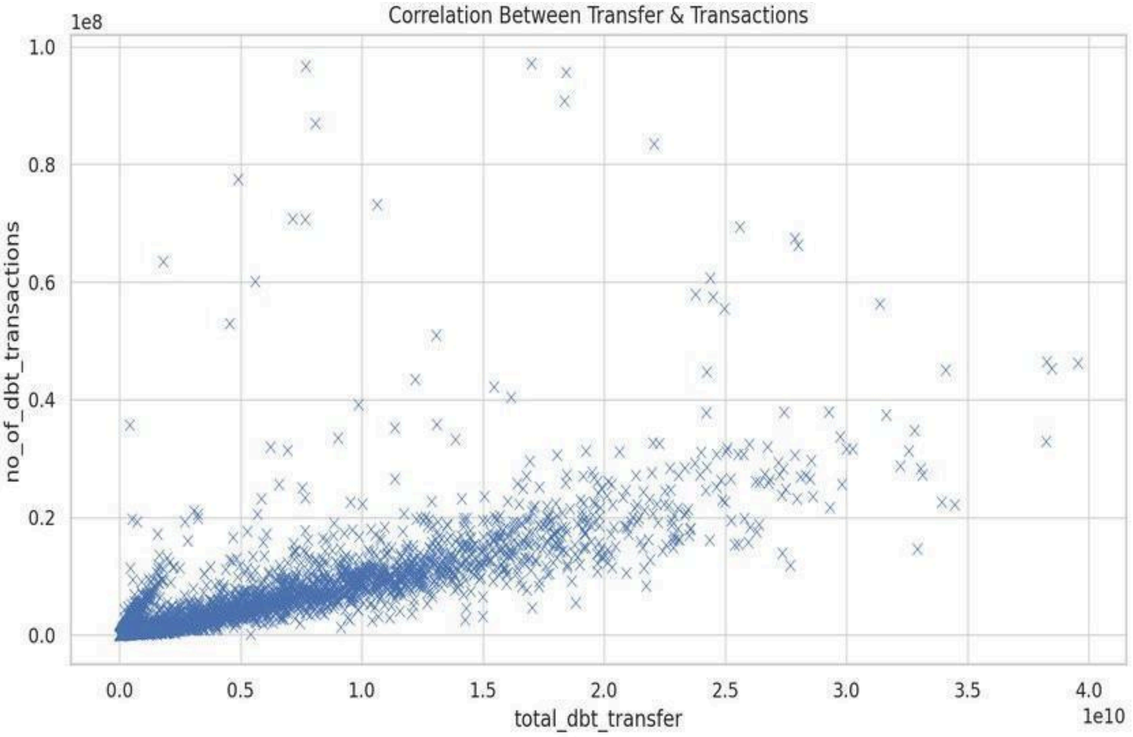
Total DBT Transactions by State

By counting the number of DBT transactions in each state, we can gauge how actively the population is engaging with subsidy or benefit schemes. High transaction volumes may indicate high coverage or better infrastructure



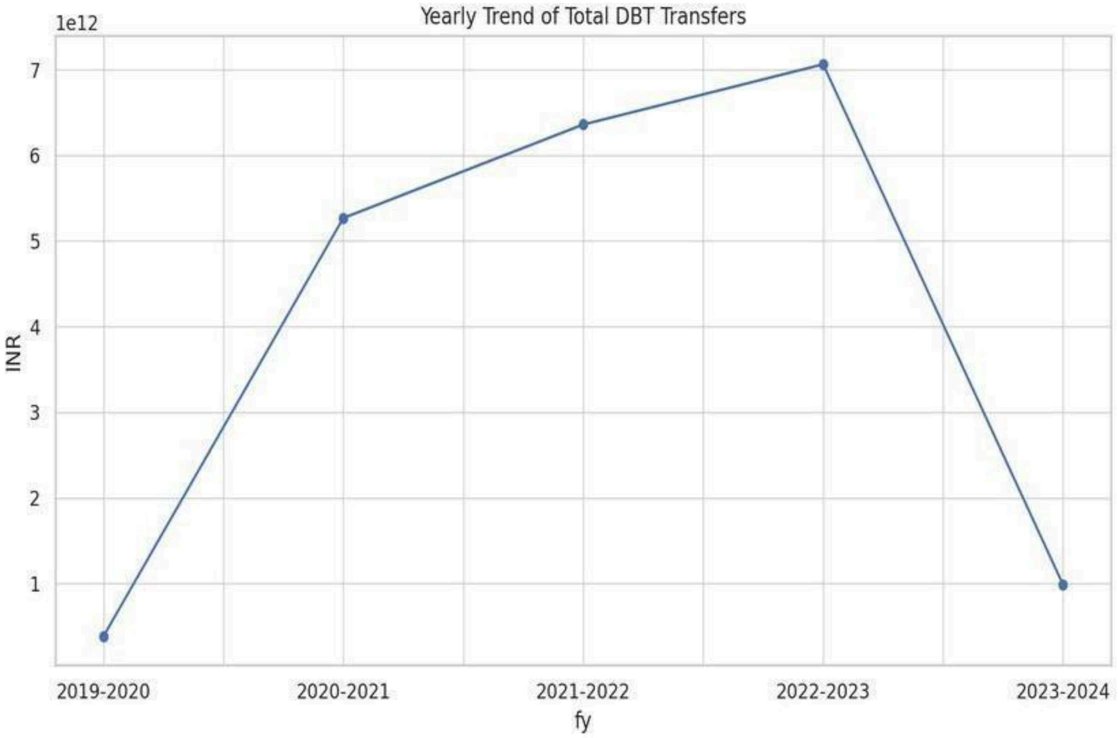
Correlation Between Transfers and Transactions

Correlation analysis identifies whether there's a statistical relationship between the number of transactions and the amount transferred. A strong positive correlation suggests that more transactions generally involve higher total transfer values.



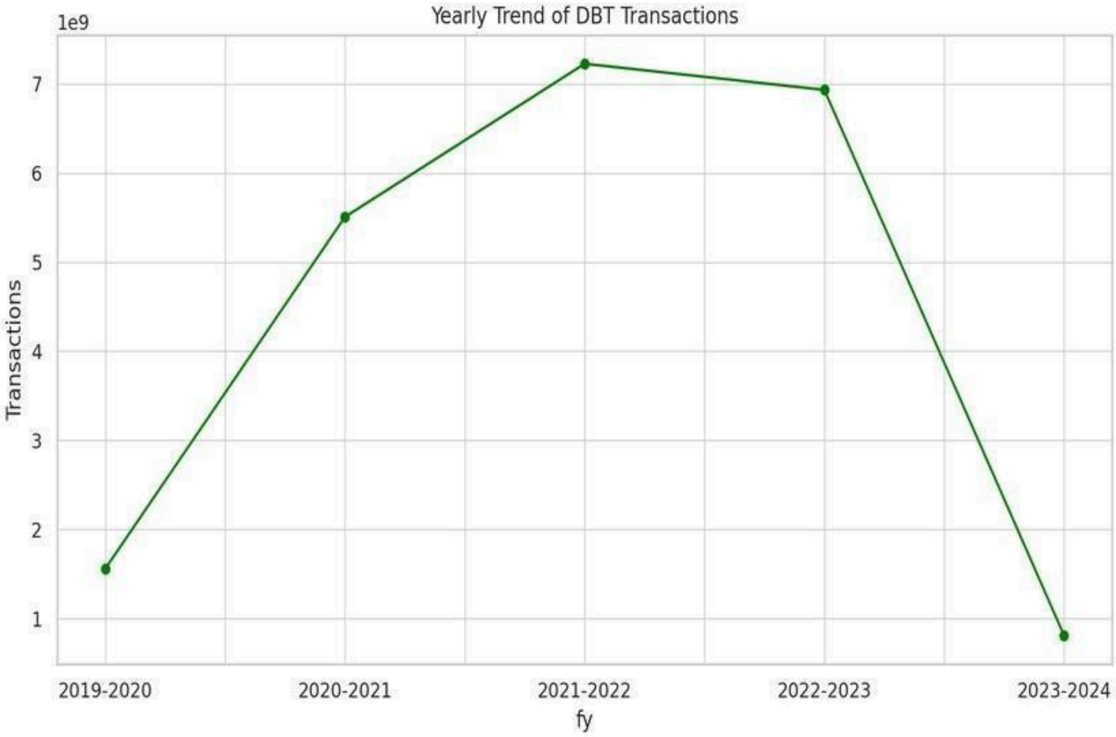
Yearly Trend of DBT Transfers

Analyzing yearly trends helps assess whether DBT transfers are increasing over time. This can reflect improvements in government outreach, digital infrastructure, or increased beneficiary participation.



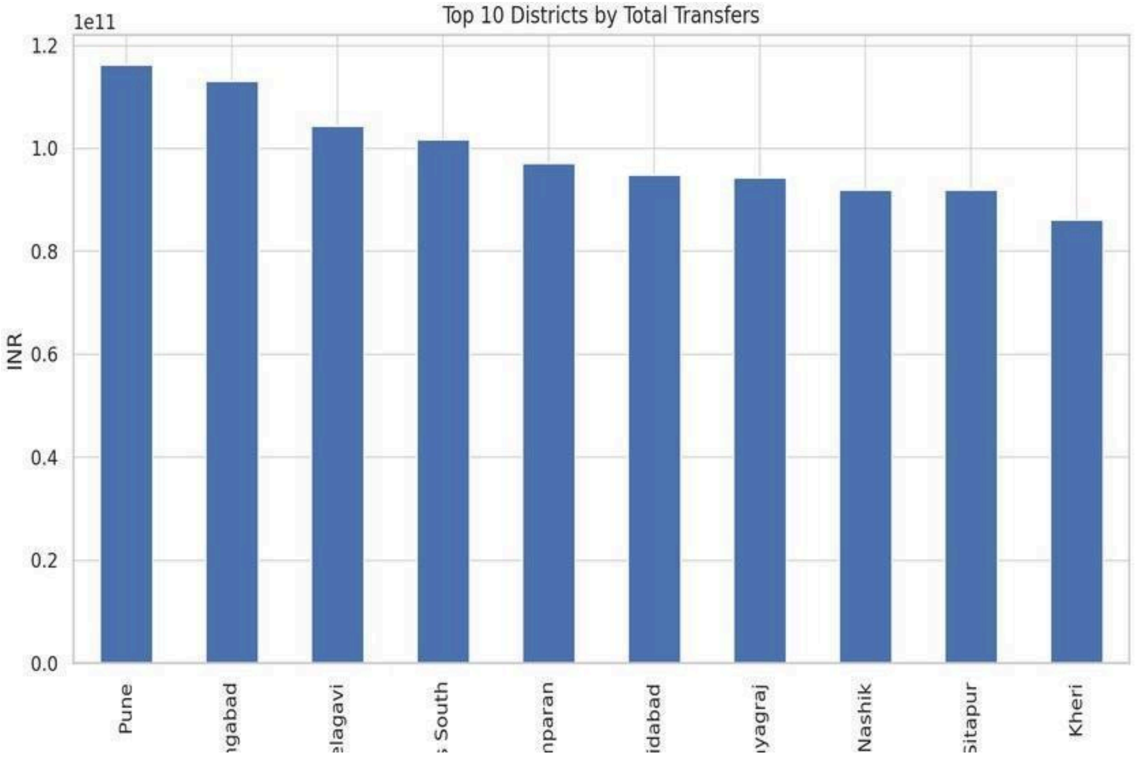
Yearly Trend of DBT Transactions

Similar to transfer trends, this analysis focuses on the number of transactions year-wise. It provides insight into growth in usage and acceptance of DBT mechanisms over time.



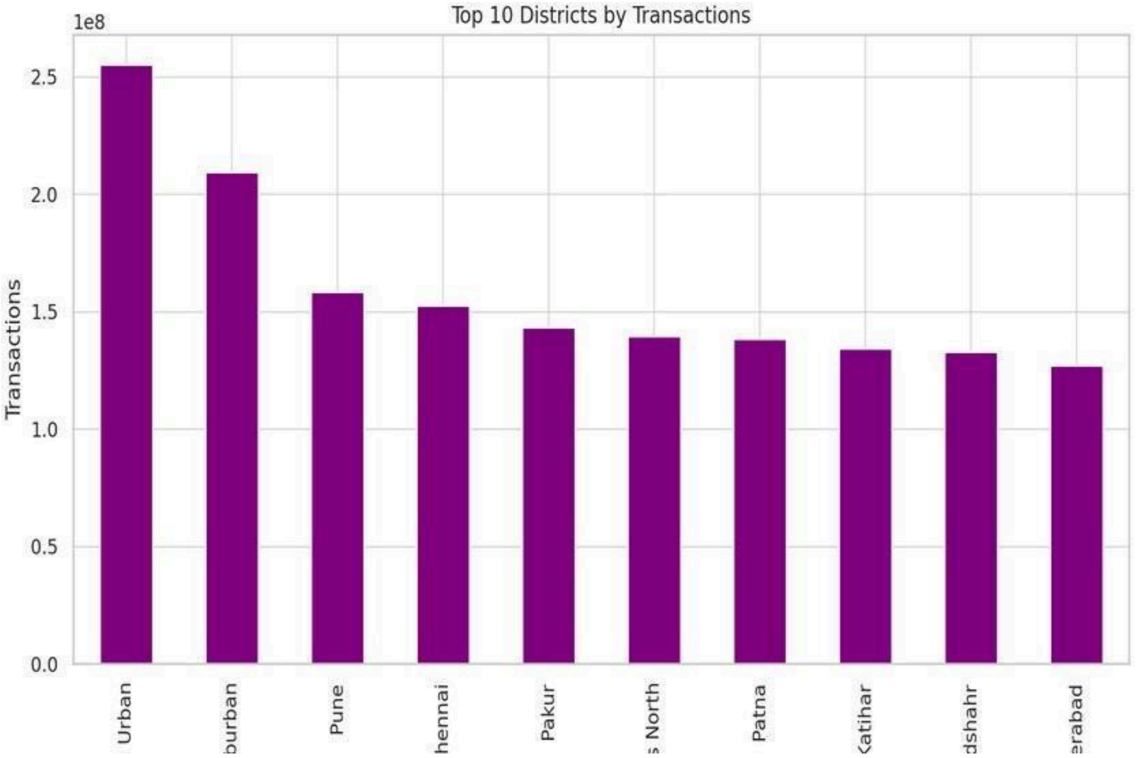
Top 10 Districts by Transfers

Identifying top-performing districts based on total transfer amount helps highlight regions with high government investment. It also uncovers spatial disparities in fund allocation.



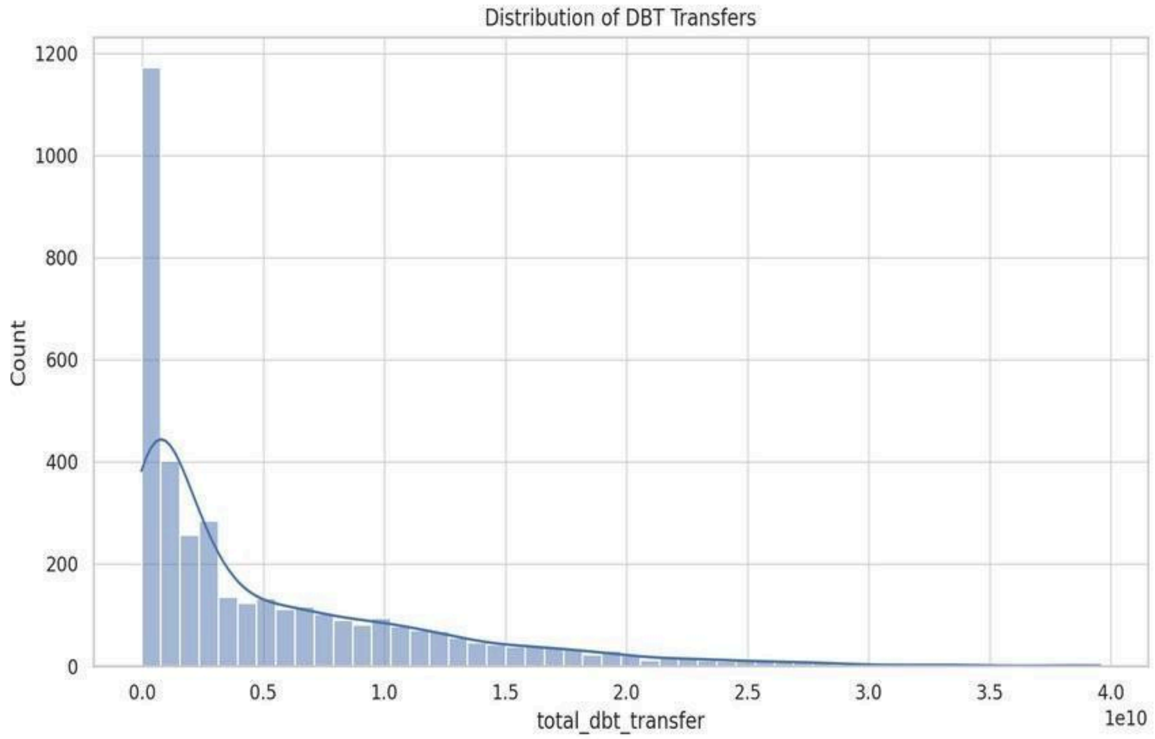
Top 10 Districts by Transactions

This highlights districts with the highest DBT activity. Frequent transactions may point to high population coverage or active participation in schemes.



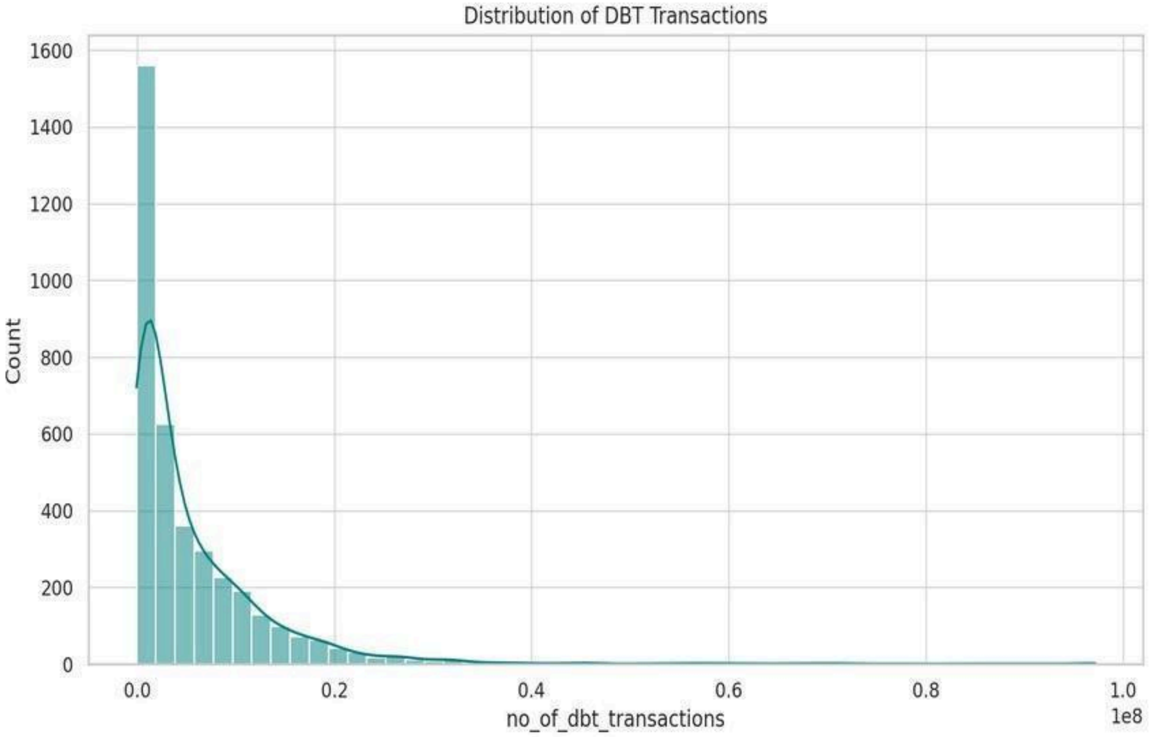
Distribution of DBT Transfers

This explores the statistical distribution (e.g., spread, skewness) of DBT amounts across all records, providing insight into whether most values are concentrated at a particular range.



Distribution of DBT Transactions

This examines how transaction counts are spread across records. It helps determine if a few districts are dominating activity or if the distribution is even.



Conclusion

This exploratory data analysis of the Direct Benefit Transfer (DBT) dataset has provided key insights into the distribution and trends of government transfers across Indian states and districts. The cleaning process ensured the dataset was free of missing values and duplicates, making the analysis reliable and consistent.

Key findings include:

- Significant variation in DBT transfers and transactions across different states and districts.
- A strong positive correlation between the number of transactions and the total amount transferred.
- An overall increasing trend in DBT activity over the years, indicating growing adoption and coverage.
- Identification of top-performing districts in terms of transaction volume and amount transferred, which could help in policy evaluation and resource allocation.