SARVA SHIKSHA ABHIYAN
EXPLORATORY DATA ANALYSIS

Course Code: IT495
Semester 2
Course Instructor: Gopinath Panda
Group Number – 01

GROUP MEMBERS:

Muskan Khare
202218037

Riya Kumari
202218049

Dhruv Solanki
202218053

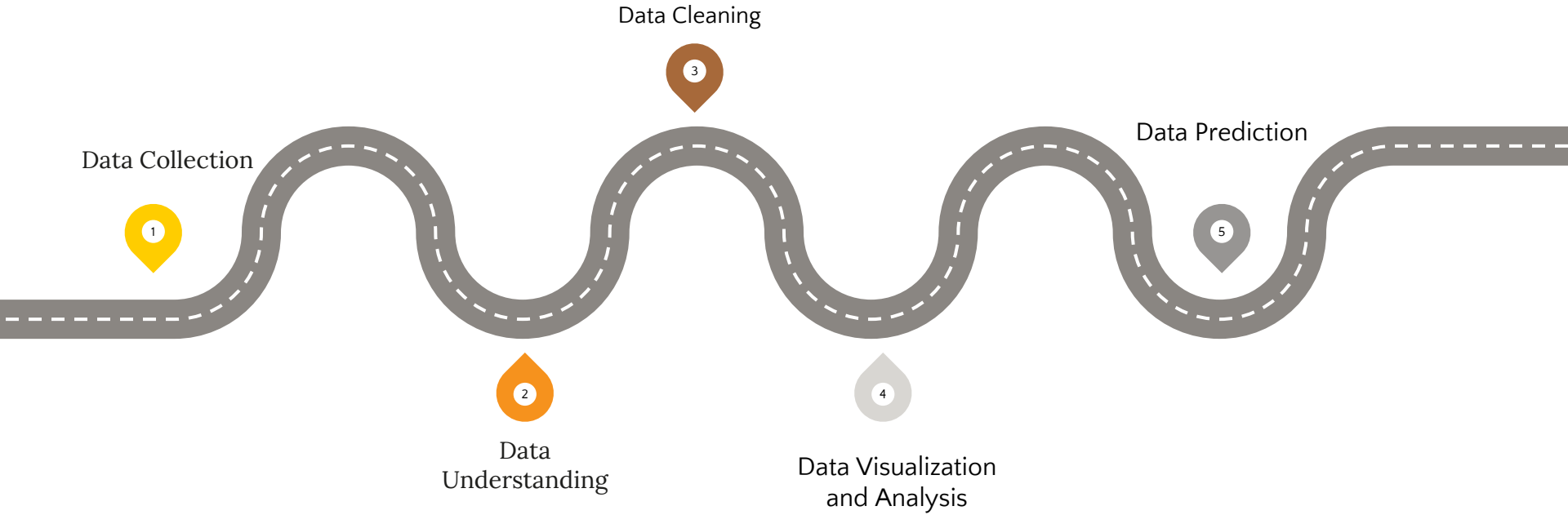
Chinmaya Pandey
202218054

Jatan Sahu
202218061

Problem Statement

- Sarva Shiksha Abhiyan (SSA) is a flagship program of the Indian government who aims at providing free and compulsory education to all children in the age group of 6 to 14 years.
- Project involves analyzing historical budget allocation and expenditure data to identify patterns and trends.
- Using this analysis, a predictive model is developed to estimate the budget allocation for the next year

5 Steps of EDA



1

Dataset Collection

- Extracted from *openbudgetsindia.org*

Dataset Description

- Dataset contains 999 rows and 26 columns.
- The dataset comprises following features:
  - State
  - Financial Year
  - Funds Released by the Government of India
  - Funds Released by the States/UTs
  - Total Funds Released (Government of India and States' Share)
  - Expenditure Incurred by the States/UTs.
  - Unspent Balance.
  - Extent of Funds Released against Budget Approved (%).
  - Extent of Funds Utilised against Budget Approved (%).

2

Dataset Understanding

- Understand the attributes of the data.
- Summarize the data by identifying key characteristics
- Understand the problems with the data

Libraries Used

- Pandas
- Numpy
- Matplotlib.pyplot
- Missingno
- Sklearn

Data Information

‘data.info()’

gives information about whole Dataset

```
ssa.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 999 entries, 0 to 998
Data columns (total 26 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   state                                   111 non-null    object
 1   code                                    111 non-null    float64
 2   year                                    111 non-null    object
 3   budget_approved                         108 non-null    float64
 4   released_funds_by_goi                   108 non-null    float64
 5   released_funds_by_states                105 non-null    float64
 6   total_funds_released                    105 non-null    float64
 7   expense_incurred_by_states              108 non-null    float64
 8   unspent_balance                         108 non-null    float64
 9   funds_released_against_budgetapproved(%)  105 non-null  float64
 10  funds_utilised_against_budgetapproved(%)  108 non-null  float64
 11  Unnamed: 11                             0 non-null      float64
 12  Unnamed: 12                             0 non-null      float64
 13  Unnamed: 13                             0 non-null      float64
 14  Unnamed: 14                             0 non-null      float64
 15  Unnamed: 15                             0 non-null      float64
 16  Unnamed: 16                             0 non-null      float64
 17  Unnamed: 17                             0 non-null      float64
 18  Unnamed: 18                             0 non-null      float64
 19  Unnamed: 19                             0 non-null      float64
 20  Unnamed: 20                             0 non-null      float64
 21  Unnamed: 21                             0 non-null      float64
 22  Unnamed: 22                             0 non-null      float64
 23  Unnamed: 23                             0 non-null      float64
 24  Unnamed: 24                             0 non-null      float64
 25  Unnamed: 25                             0 non-null      float64
dtypes: float64(24), object(2)
```

Descriptive Analysis

	budget_approved	released_funds_by_goi	released_funds_by_states	total_funds_released	expense_incurred_by_states	unspent_balance
count	108.000000	108.000000	105.000000	105.000000	108.000000	108.00000
mean	2027.298472	609.993204	598.864952	1213.097905	1295.056259	112.89813
std	3424.095576	898.557443	1415.677913	2274.724619	2257.073192	240.84146
min	3.118000	0.784000	0.000000	0.784000	2.305000	-364.90800
25%	195.194750	80.341500	10.794000	92.595000	123.583000	2.93350
50%	953.287500	311.795000	110.855000	408.910000	535.006000	32.50450
75%	2323.564000	777.405000	518.270000	1305.036000	1453.842250	128.23375
max	20688.135000	5043.183000	9404.330000	14447.513000	14588.360000	1195.14400

8 rows × 23 columns

Summary Of the Data (including 5 number summary)

3

Dataset Cleaning

- Handling missing values.

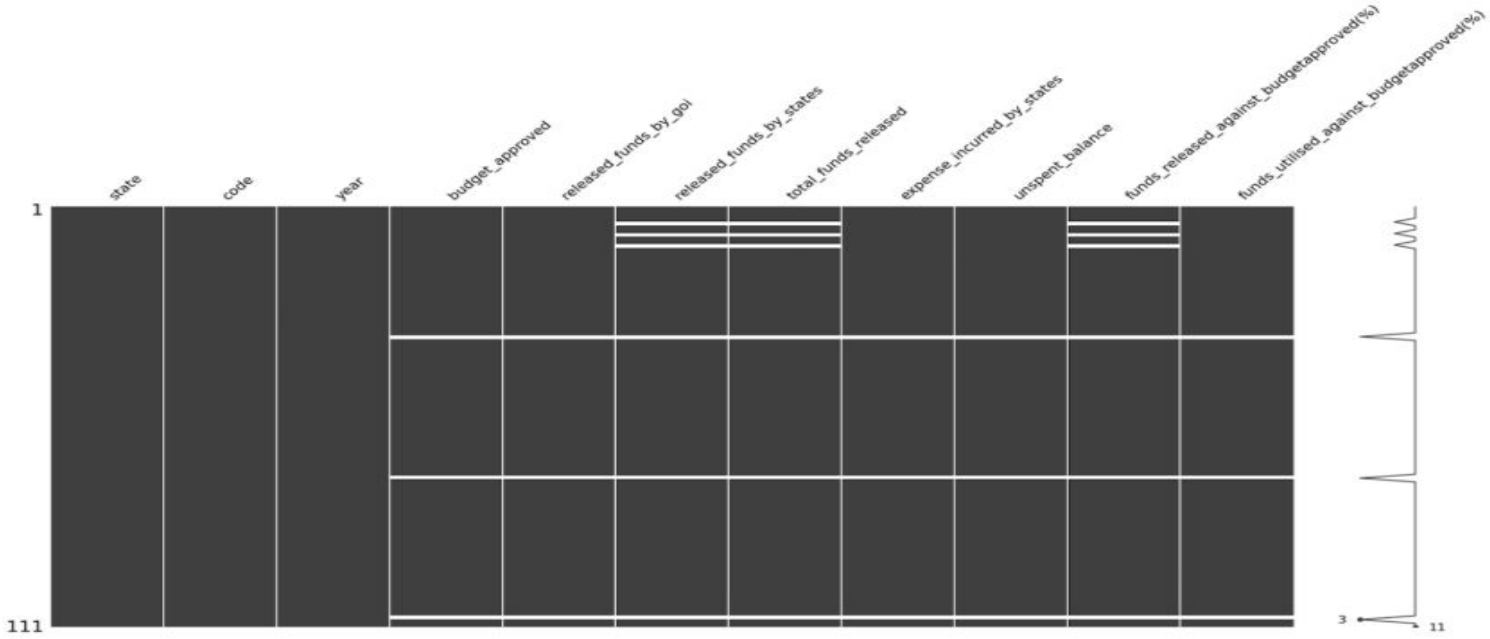
Checking Missing Value

- Checking for missing values in data is a crucial step in data analysis and preprocessing.
- Missing values refer to the absence of data entries or values in specific variables or observations within a dataset.
- Checking for missing values is important because it can affect the accuracy and reliability of any data analysis.

```
#Checking null values (NUMERICAL)
ssa.isna().sum()
```

```
state                                    888
code                                     888
year                                     888
budget_approved                          891
released_funds_by_goi                    891
released_funds_by_states                 894
total_funds_released                     894
expense_incurred_by_states               891
unspent_balance                          891
funds_released_against_budgetapproved(%) 894
funds_utilised_against_budgetapproved(%) 891
Unnamed: 11                              999
Unnamed: 12                              999
Unnamed: 13                              999
Unnamed: 14                              999
Unnamed: 15                              999
Unnamed: 16                              999
Unnamed: 17                              999
Unnamed: 18                              999
Unnamed: 19                              999
Unnamed: 20                              999
Unnamed: 21                              999
Unnamed: 22                              999
Unnamed: 23                              999
Unnamed: 24                              999
Unnamed: 25                              999
dtype: int64
```

Visualizing Missing Values



Matrix plot visualization of missing values using msno.matrix(ssa)

Drop/Delete Missing Values

	state	code	year	budget_approved	released_funds_by_goi	released_funds_by_states	total_funds_released	expense_incurred_by_states	unspent_balance
34	Ladakh	35.0	2015-2016	NaN	NaN	NaN	NaN	NaN	NaN
71	Ladakh	35.0	2016-2017	NaN	NaN	NaN	NaN	NaN	NaN
108	Ladakh	35.0	2017-2018	NaN	NaN	NaN	NaN	NaN	NaN

We can see that the ‘budget_approved’ column is null for ‘state’ Ladakh. Furthermore, all columns of Ladakh are null. Since this data won't be beneficial for our analysis it's better to delete this from our dataset.

Other Null Values

```
# Rows which have null values
ssa_row= ssa[ssa.isna().any(axis=1)]
ssa_row
```

	state	code	year	budget_approved	released_funds_by_goi	released_funds_by_states	total_funds_released	expense_incurred_by_states	unspent_balance
4	Chhattisgarh	5.0	2015-2016	2149.343	622.197	NaN	NaN	1477.519	36.023
7	Haryana	8.0	2015-2016	1120.583	345.012	NaN	NaN	529.163	99.413
10	Karnataka	11.0	2015-2016	1545.808	417.593	NaN	NaN	1196.365	38.156

For Chhattisgarh, Haryana and Karnataka we have null values in some cells. We need to handle these missing values in order to further work on our data.

Types of Missing Values

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

Here the missingness depends on other observed variables in the dataset, but not on the missing variable itself. Hence our data is **Missing at Random(MAR)**, so we impute it by using their respective mean and pattern.

Imputation

- For column “released_funds_by_states” we have filled the null values by taking the mean of other years’ data.

- For column “total_funds_released” we have filled null values using the formula below -

  total_funds_released = released_funds_by_goi + released_funds_by_states

- For column “funds_released_against_budget_approved” we have filled null values using the formula below -

  funds_released_against_budget_approved = (total_funds_released/budget_approved)×100

Verifying Null Values

After completing the imputation process, it is essential to conduct a thorough verification of null values to ensure that no additional missing values have been overlooked.

```
ssa.isnull().sum()
```

```
state                                    0
code                                     0
year                                     0
budget_approved                          0
released_funds_by_goi                    0
released_funds_by_states                 0
total_funds_released                     0
expense_incurred_by_states               0
unspent_balance                          0
funds_released_against_budgetapproved(%) 0
funds_utilised_against_budgetapproved(%) 0
dtype: int64
```

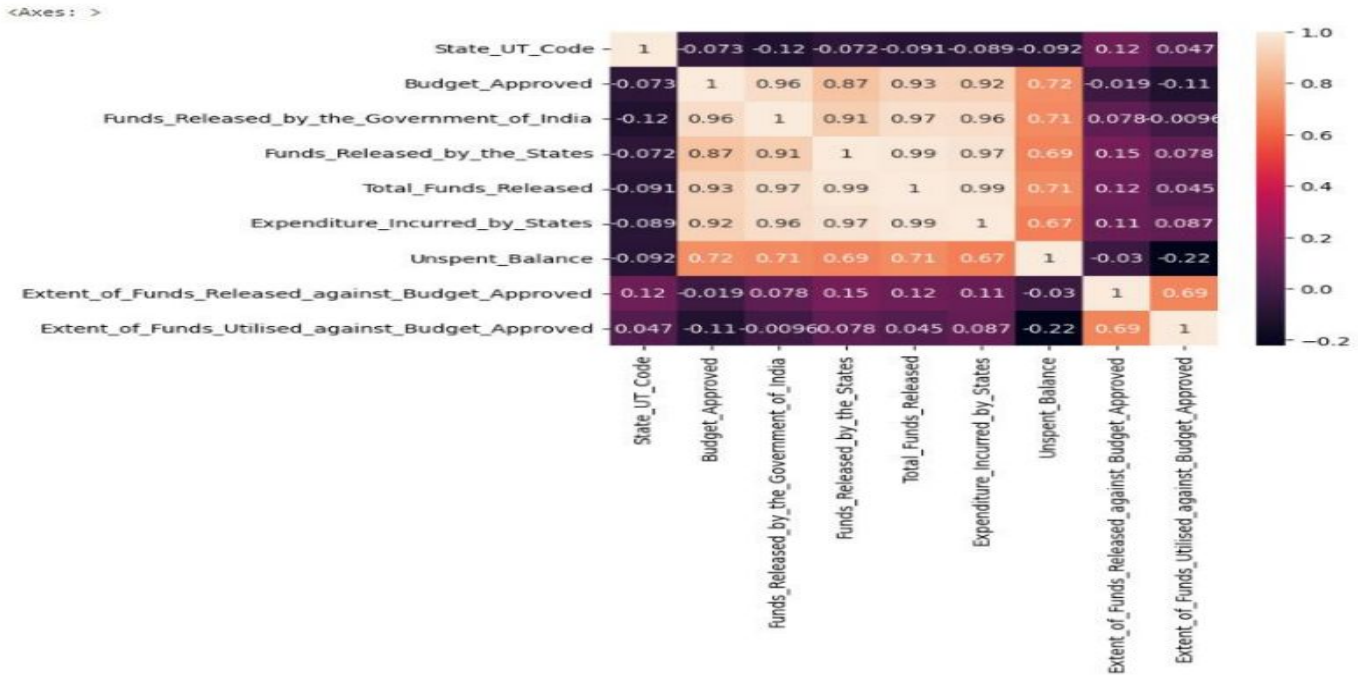
4

Dataset Visualization and Analysis

- Representation of Data through use of common Graphics.
- Easy way to convey concepts in a universal manner.

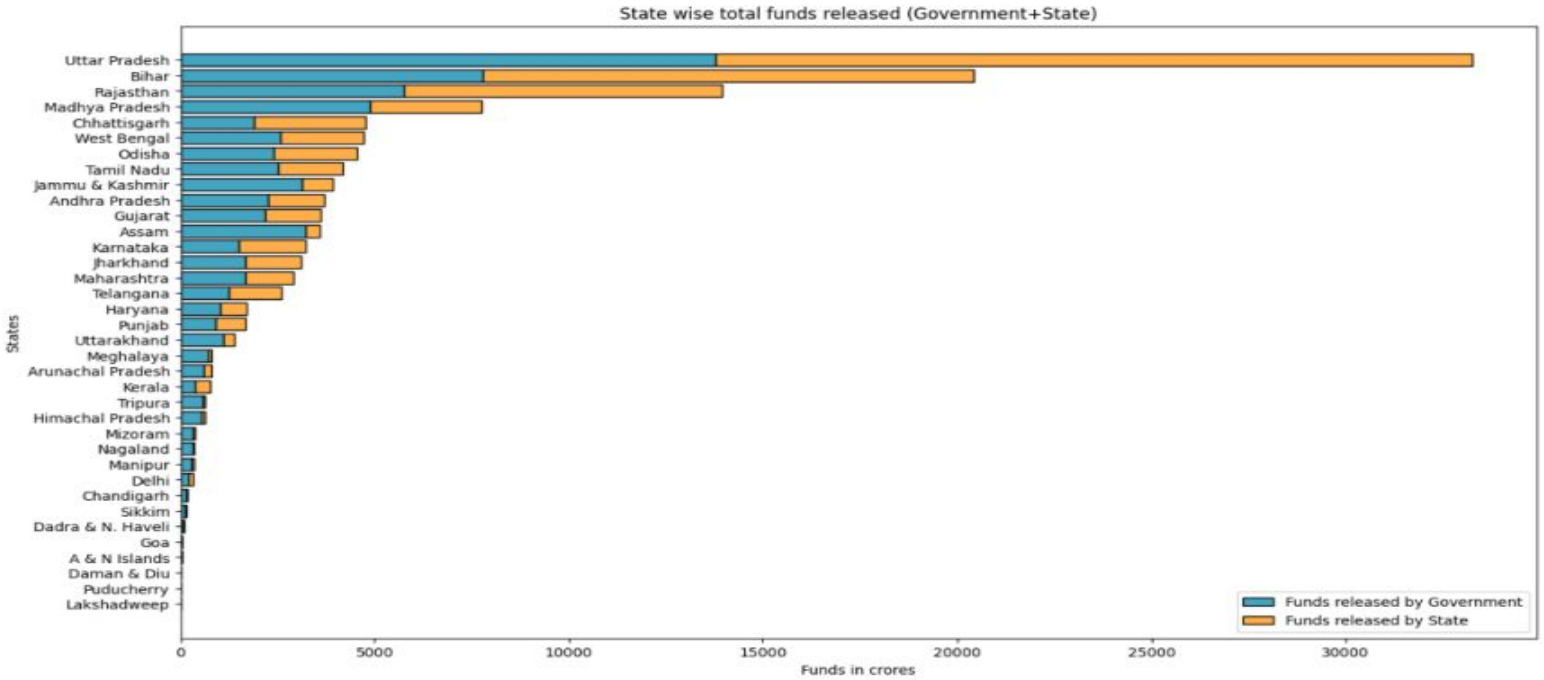
Heatmap

Correlation Map using Seaborn Library

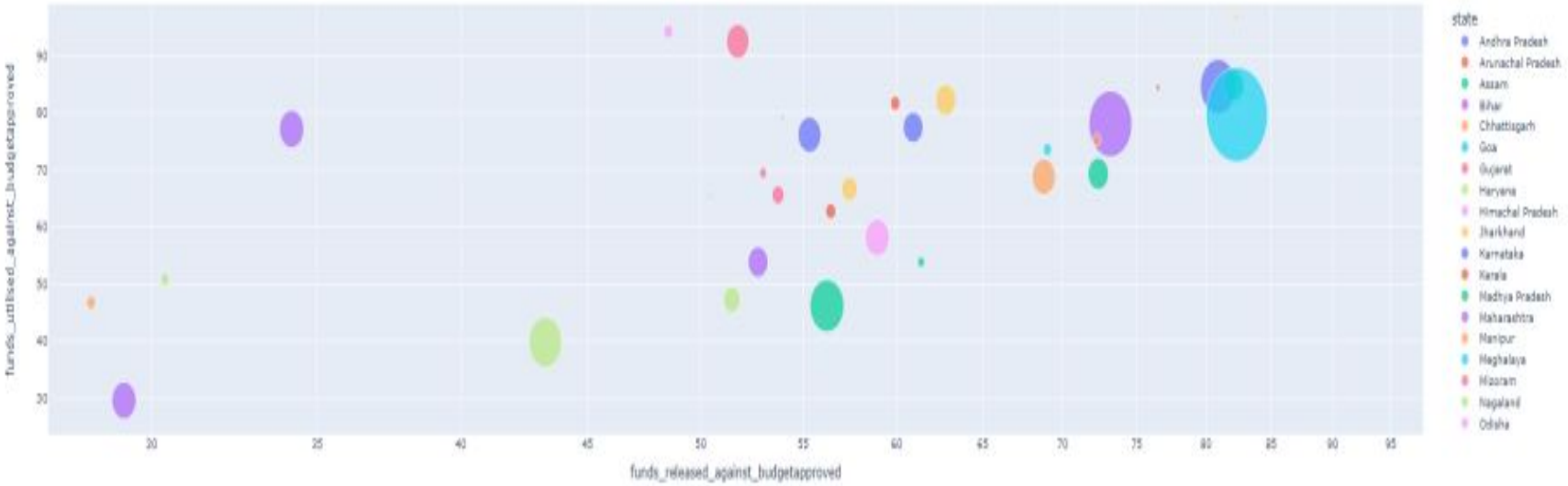


Stacked Bar Chart

Distribution of the funds among each state. (Using Matplotlib)



Weighted Scatter Plot

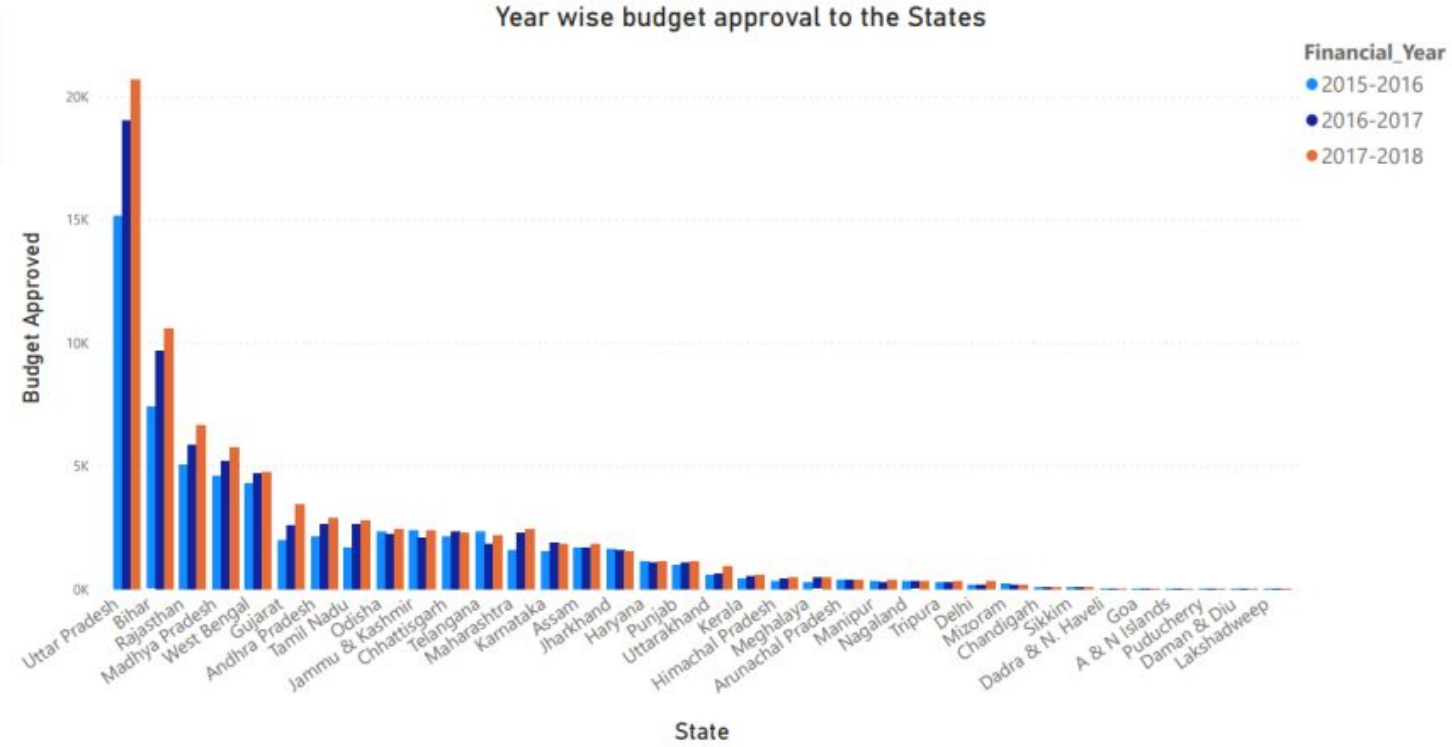


Scatter plot is of funds released vs utilised against the budget. Weights in the scatter is of the budget approved. We can see that as funds released increases the utilization of that funds also increases.

Visualization Dashboard

- A visualization dashboard is a collection of visualizations and interactive components that provide a high-level overview of key performance indicators (KPIs).
- Visualization dashboards can be created using various tools and software, including Excel, Tableau, Power BI, and other data visualization software.
- Here, we have used Power BI to make the dashboard. We have made static visuals and dynamic dashboard for the project.

Static Visual



Dynamic Dashboard

A dynamic dashboard is an interactive and customizable data visualization tool that provides real-time updates and allows users to explore and analyze data dynamically. It typically consists of multiple visual components, such as charts, graphs, tables, and filters, that are interconnected and update automatically as the underlying data changes.

Power Bi Dashboard - Link

5

Dataset Prediction

- Feature Selection
- Machine Learning Modeling

Feature Selection

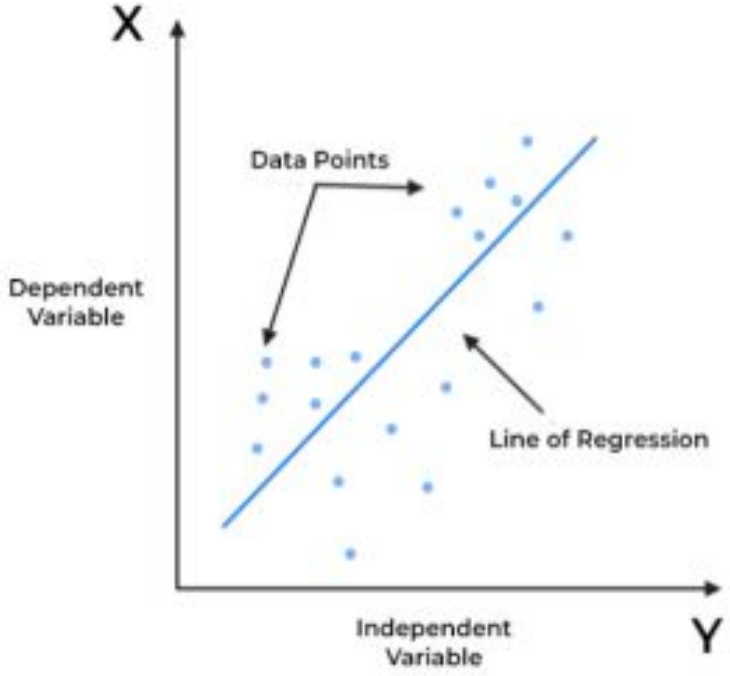
- Finding Correlation
  - Taking state_code and year as independent variable.
  - Taking approved budget as dependent variable.
- Preprocessing Data For Modelling
  - Encoding year.
  - Creating dataframes for specific year.

Machine Learning Modelling

- ***Objectives***
  - ➢ OBJECTIVE 1- We will train data of 2 years(2015-16 to 2016-17) and will take 'approved_budget' as test data for year 2017-18 and decide which ML model will work efficiently.
  - ➢ OBJECTIVE 2 - We will take ML model from objective 1 for better result and train data of 3 years(2015-16 to 2017-18) and hence will predict 'budget_approved' for 2018-19 financial year.

Objective 1

- **Using Linear Regression**
  - ➢ Linear regression is a statistical method used to analyze the relationship between a dependent variable and one or more independent variables.
  - ➢ The goal of linear regression is to find the linear relationship between the variables, which can be used to make predictions.

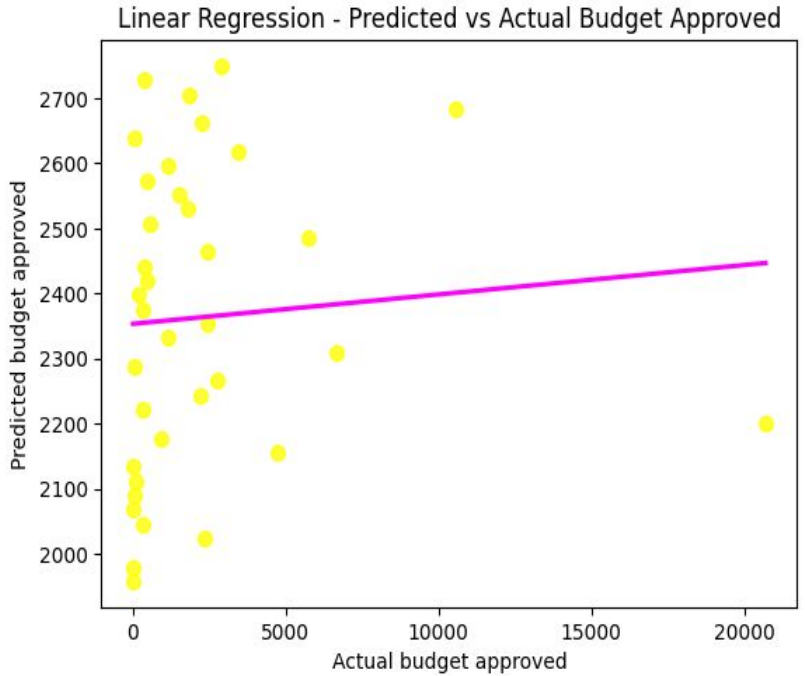


Objective 1

USING LINEAR REGRESSION

- Data is non linear so it will not be a good model to use linear data.

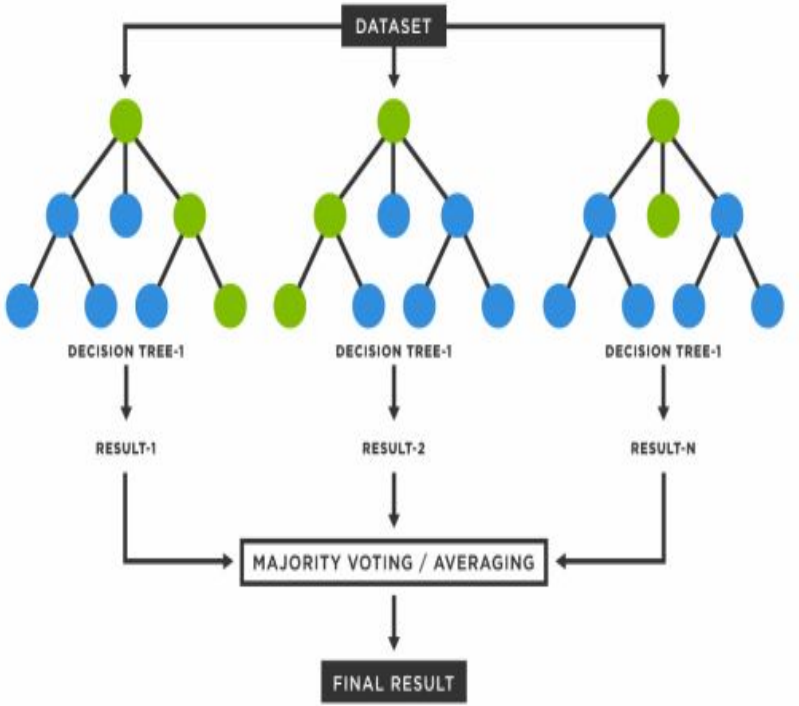
```
Mean absolute error:   2239.904197563468
```



Objective 1

USING RANDOM FOREST

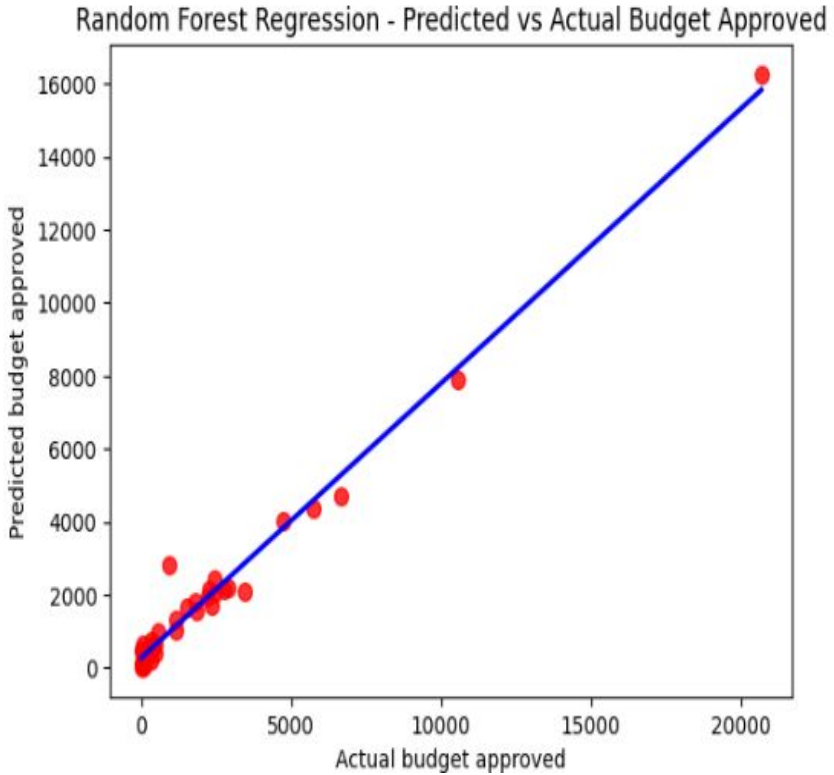
- Random forest is a machine learning algorithm that is used for classification, regression, and other tasks that involve supervised learning
- It is an ensemble method that combines multiple decision trees to improve accuracy and reduce overfitting.

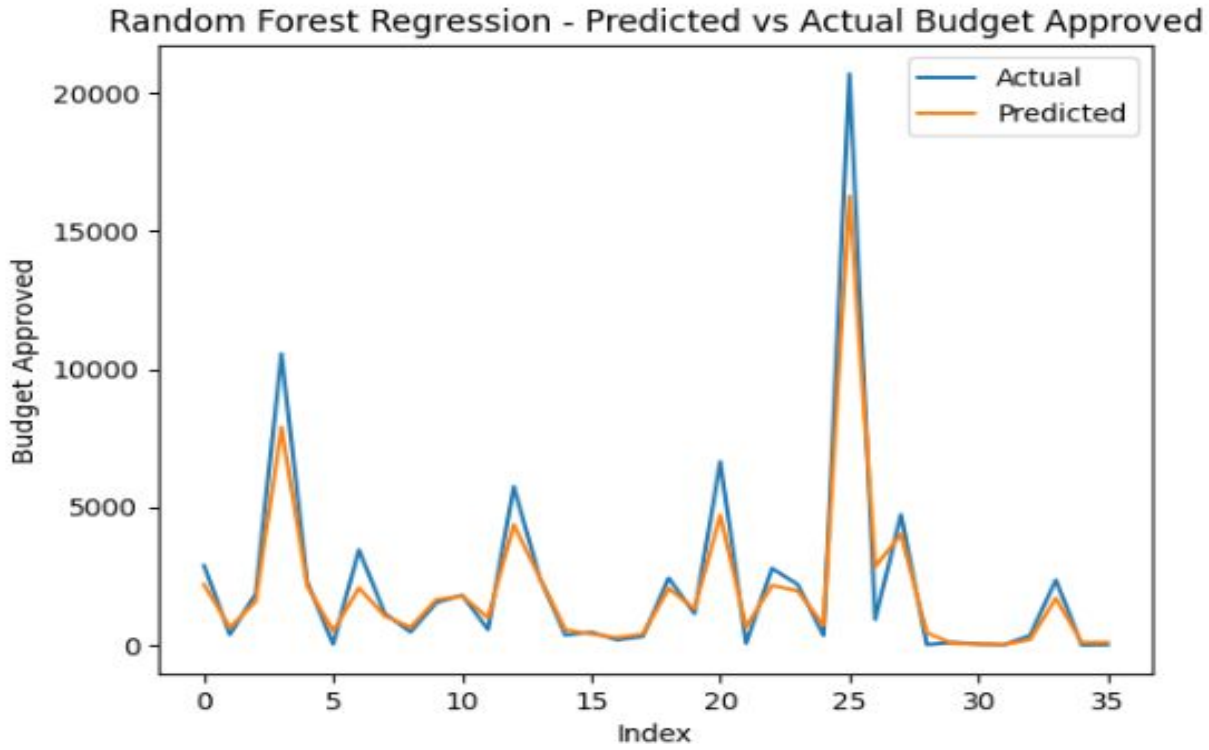


Objective 2

- **USING RANDOM FOREST**

  - After applying we find that our accuracy is increased and error is decreased.





Mean Absolute Error: 592.4466100000001

Objective 2

RESULT :-

- Prediction for approved_budget for 2018-19

  - ➢ Due to availability of limited data to predict budget for next year our mean absolute error is high.

```
Mean Absolute Error:  592.4466100000001
```

	state	approved_budget
0	Andhra Pradesh	2544.05732
1	Arunachal Pradesh	585.52034
2	Assam	1729.43079
3	Bihar	9364.90411
4	Chhattisgarh	2411.12661
5	Goa	585.65834
6	Gujarat	2738.89702
7	Haryana	1049.17183
8	Himachal Pradesh	601.97063
9	Jharkhand	1514.36460
10	Karnataka	1754.81993
11	Kerala	864.99699
12	Madhya Pradesh	5337.79412
13	Maharashtra	2782.07444
14	Manipur	544.78667
15	Meghalaya	412.42726
16	Mizoram	239.54905
17	Nagaland	323.39707

Thank You!