DAV TEAM ASSIGNMENT

By : Varun Ram Narayanan 22B0347

Table of Contents

Question Number	Description
1. NDAP Data Analysis	Analyzing India’s energy landscape data for strategic decision taking
2. Pollution Trends	Correlation of power plant data with pollution trends and identifying additonal sources.
3. Market Entry Strategy	Replication of segmented outreach strategy for 2500 potential New Customers
4. Unsupervised ML	Different Unsupervised Machine Learning Algorithms
5. Customer Segmentation	Customer Segmentation using Unsupervised ML Algorithms
6. Model Validation	Improve model accuracy for predicting renewable energy production.
7. Spam Classifier	Creating a Model to Classify Emails into Spam and Non-Spam

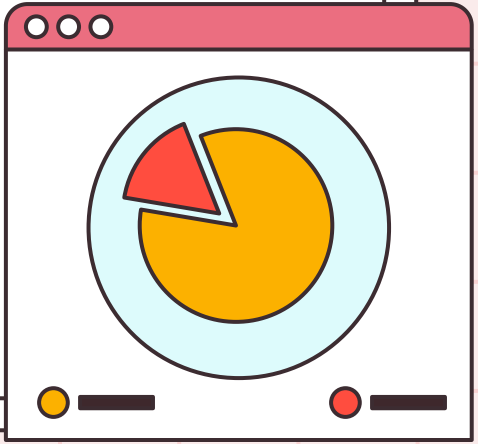
Introduction

- We have been given a dataset regarding rate of generation of energy in power plants located in India.
- The dataset has 46675 rows and 20 columns. It is a clean dataset with NaN values present only in few rows which we can drop before beginning our analysis as the number of rows getting dropped is negligible with respect to the size of the dataset.
- We must create a new column with the Actual Energy Generated in MW and similarly for Generation Programme for CEA as well to maintain unit consistency. They are initially in GW h.

01

NDAP DATA ANALYSIS

Dive into the Data: 10 Essential Questions Answered!



Part I

(a)How has the total amount of energy generated across the country varied as the year have gone by?

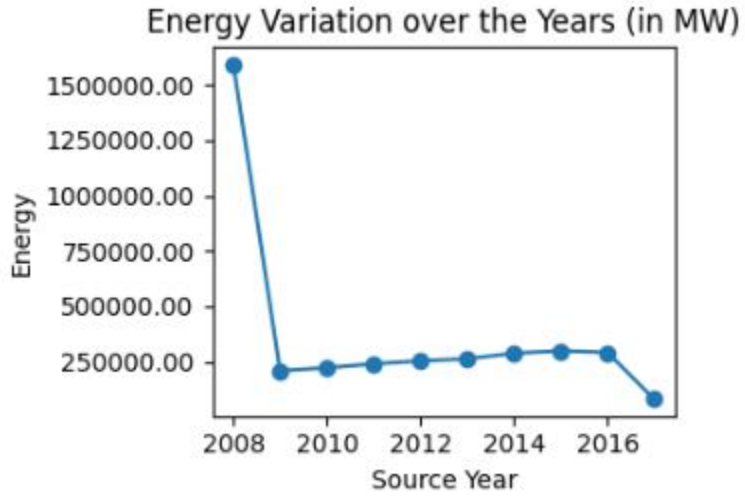
Solution :

- We will require the columns **Actual energy generated(MW),'YearCode, and 'SourceMonth'** columns.

- We will group the dataset by YearCode once and sum the total energy generated per state. Similarly, we repeat the same for SourceMonth and plot them.

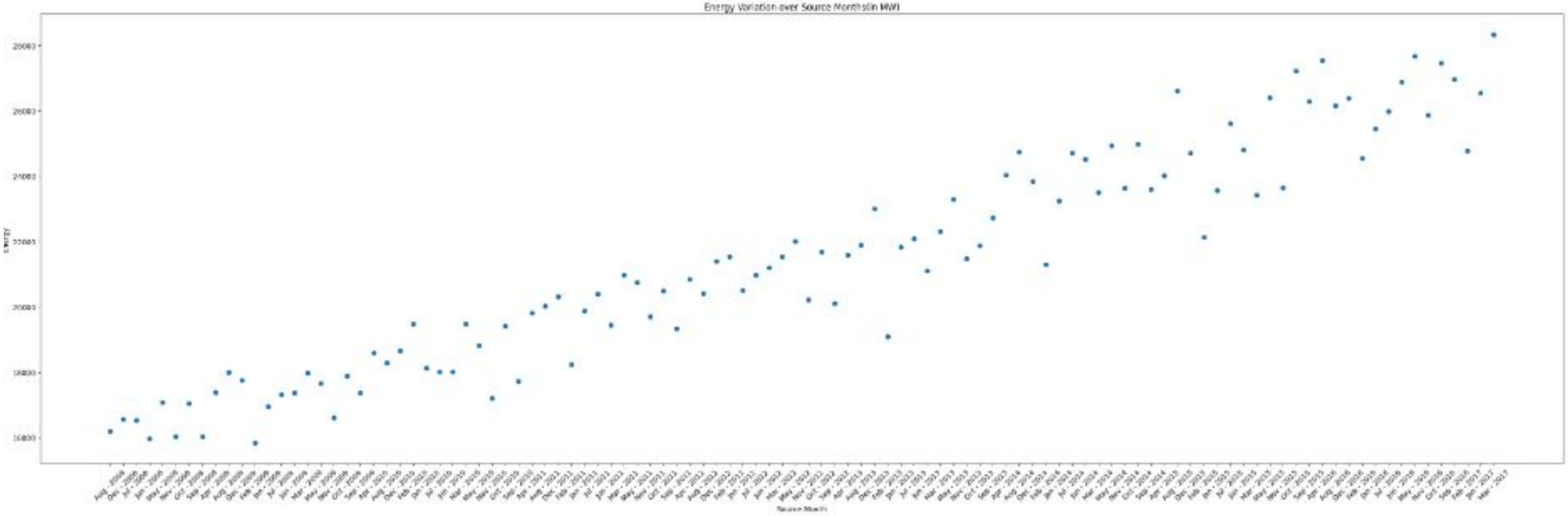
- **While plotting the time variation for SourceMonth, I have removed the energy generated in April 2008 as it was abnormally high compared to the remaining ans was skewing the plot.**

Plots



- Our first observation is that there is something anomalous about the year 2008 and amount of energy generated.

Plots



- The amount of energy generated is increasing as the years go by which is a sign of technological advancements and proper renovation of plants.

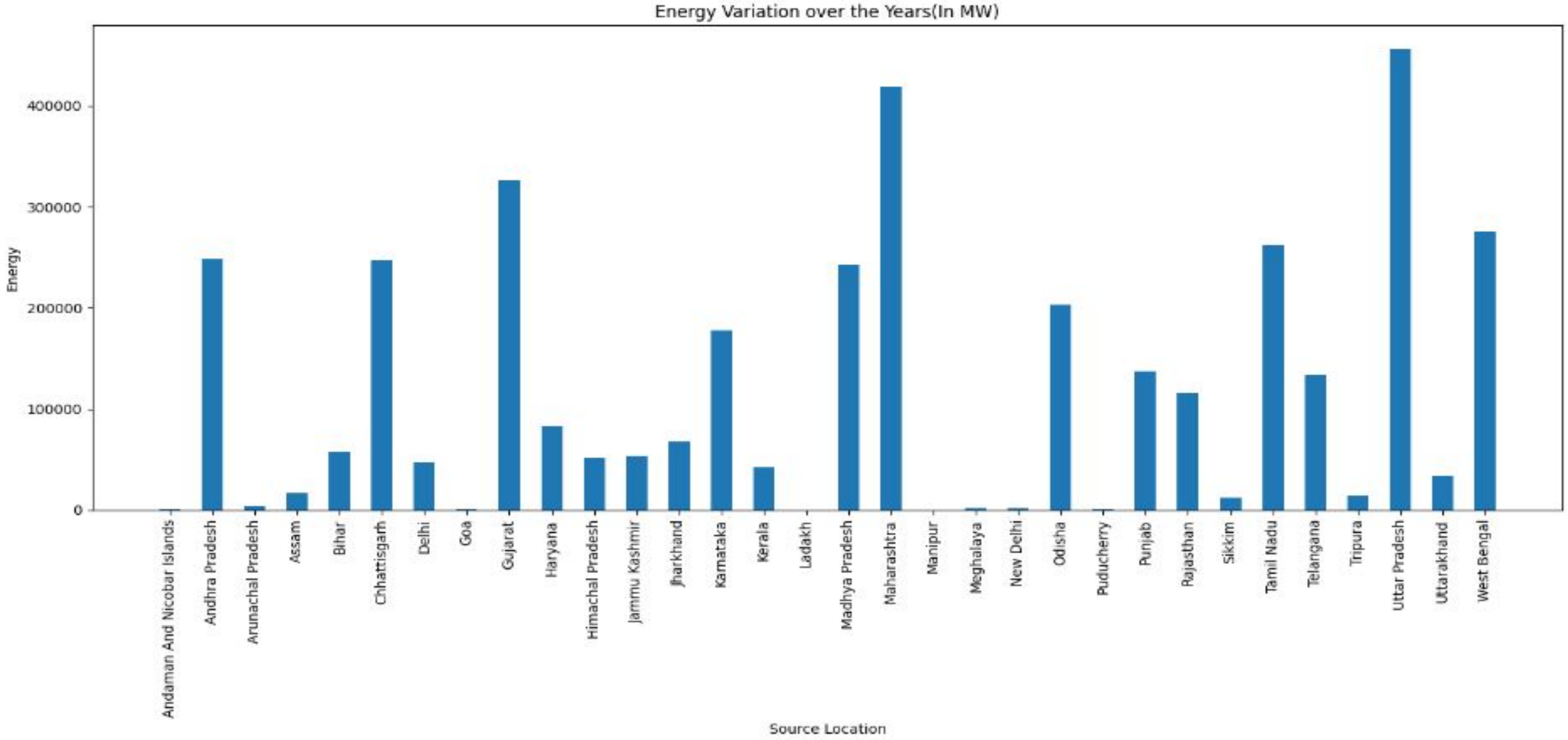
Part II

(e) What is the distribution of amount of energy generated across the country?

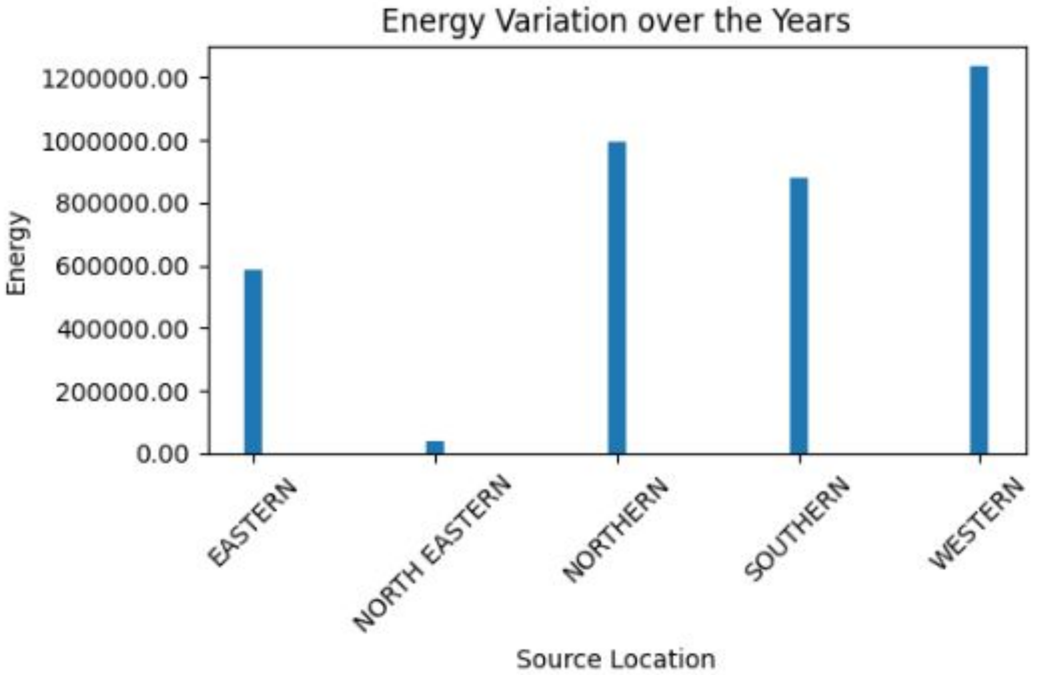
Solution :

- We will require the columns **Actual energy generated(MW),’State’, and ‘Region’** columns.
- We will group the dataset by State once and sum the total energy generated per state. Similarly, we repeat the same for Region and plot them.
- Maximum amount of energy was generated in the states of Maharashtra and Uttar Pradesh.

Plots


Energy Variation over the Years(In MW)
Energy
400000
300000
200000
100000
0
Andaman And Nicobar Islands
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chhattisgarh
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu Kashmir
Jharkhand
Karnataka
Kerala
Ladakh
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
New Delhi
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Telangana
Tripura
Uttar Pradesh
Uttarakhand
West Bengal
Source Location

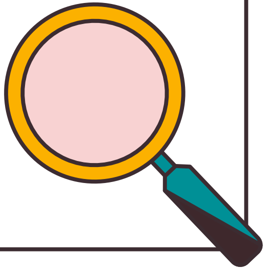
Plots



Energy Variation over the Years
1200000.00
1000000.00
800000.00
600000.00
400000.00
200000.00
0.00
Energy
EASTERN
NORTH EASTERN
NORTHERN
SOUTHERN
WESTERN
Source Location

Initial Outlook

- After looking at the first 2 questions, we can make the observation that there is a significant anomaly in the month of April 2008. Similarly the amount of energy produced in some states is quite high which is suspicious as well. This anomaly is majorly concentrated in 2008.
- This pushed me towards investigating the Installed Capacity with respect to the Actual Energy Generated in MW.
- Exceeding the Installed Capacity in a plant is quite dangerous and ill-advised. It is important to track it which is why I have separately analysed this and incorporated it into most of the remaining questions as well



Part III

(e)What percentage of plants are operating above capacity? What is the distribution of these plants across the country?

Solution :

- We form the **Difference** column by taking the difference of values in the column **'Installed Capacity' and 'Actual energy generated(MW)'.**

- By plotting the difference column, I got an idea of the distribution of values. There were positive, negative, and zero values randomly distributed.

- Plants operating above capacity was particularly suspicious and possibly unsafe.

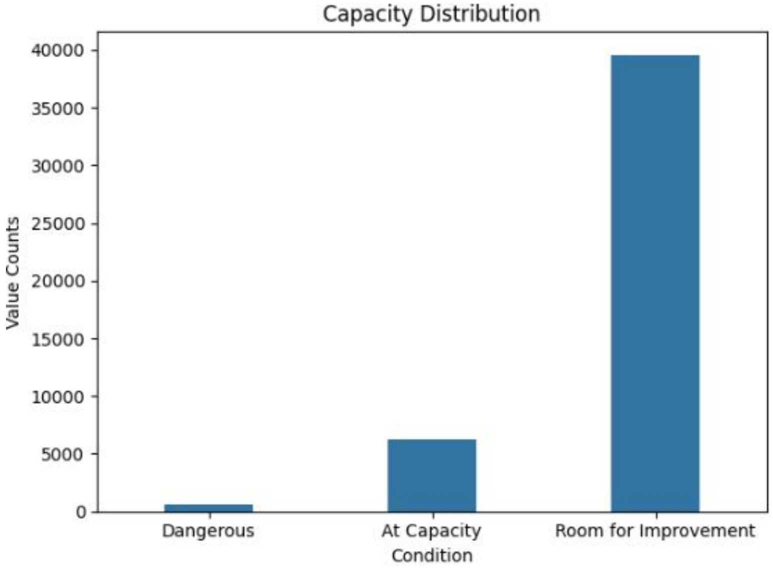
- I plotted the state-wise distribution of this Difference column as well to understand whether there was any particular concentration of these unsafe plants in specific states.

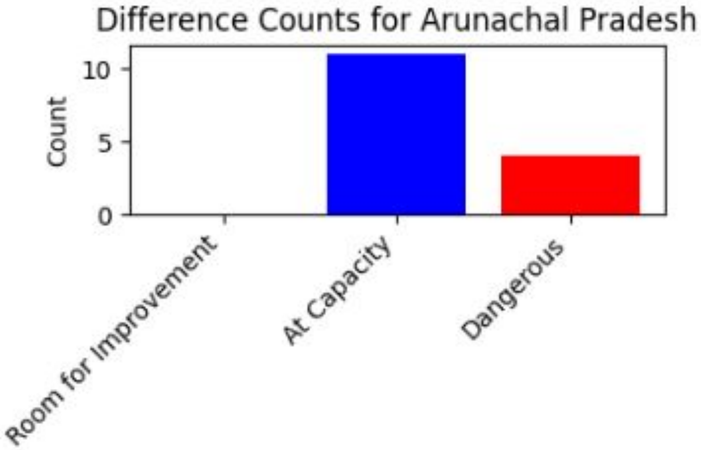
Part III (cont.)

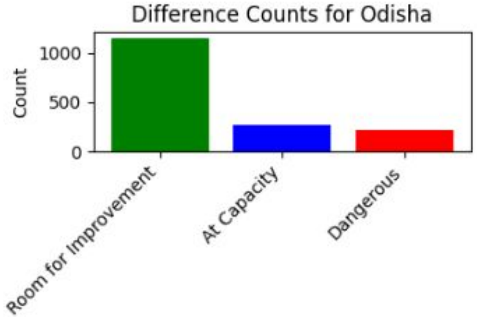
Solution(cont.) :

- **Room for improvement** implies that the plant is operating below its installed capacity.
- **At Capacity** implies that the plant is operating at its installed capacity.
- **Dangerous** implies that the plant is operating above its installed capacity which is potentially unsafe and must be tracked.
- There are a total of 603 plants operating above installed capacity which is around 1% of the total numbef of existing plants

Plots


Capacity Distribution
Value Counts
40000
35000
30000
25000
20000
15000
10000
5000
0
Dangerous
At Capacity
Room for Improvement
Condition


Difference Counts for Arunachal Pradesh
Count
10
5
0
Room for Improvement
At Capacity
Dangerous


Difference Counts for Odisha
Count
1000
500
0
Room for Improvement
At Capacity
Dangerous

Conclusions

1.

From this analysis, we can conclude **78%** of the plants are actually operating below the installed capacity. Thus, there is definitely scope to increase energy production in India in a safe manner.

2.

- There are **132** plants in Gujarat and **221** in Odisha operating above installed capacity out of the total 603. This indicates that the plants in these states need to particularly checked and made to operate at capacity for everyone’s safety

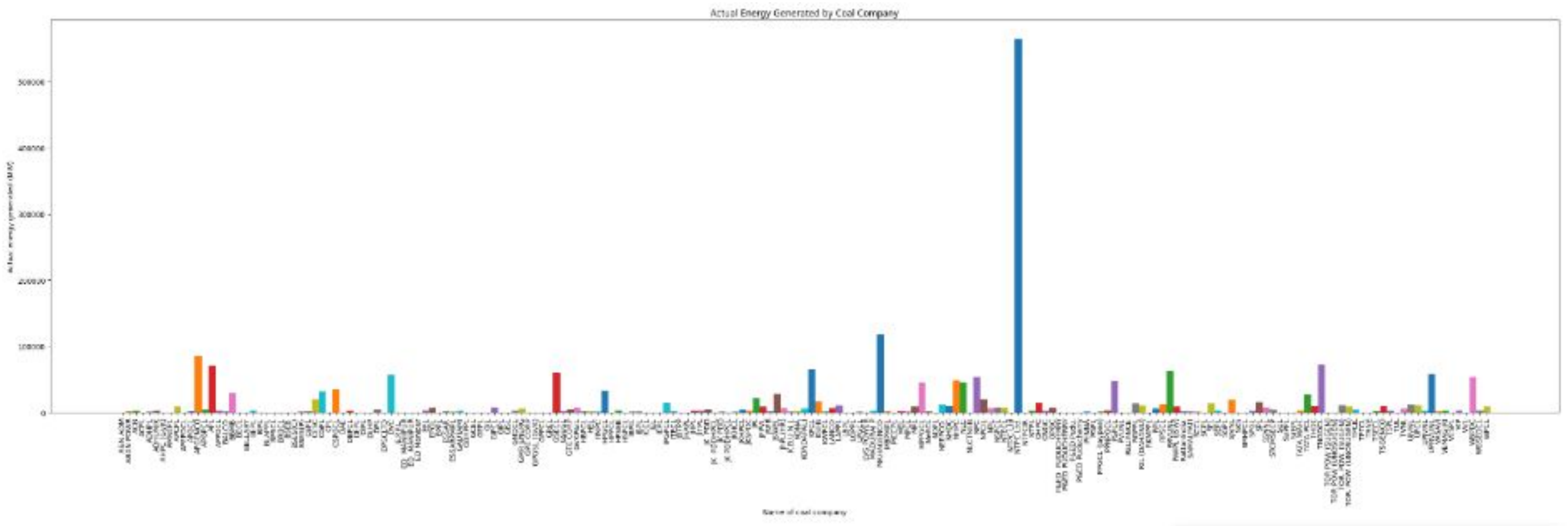
Part IV

(e) (d) What is the distribution of amount of energy generated by different coal companies? Are the plants producing this energy in a safe manner, or are they operating above capacity.?

Solution :

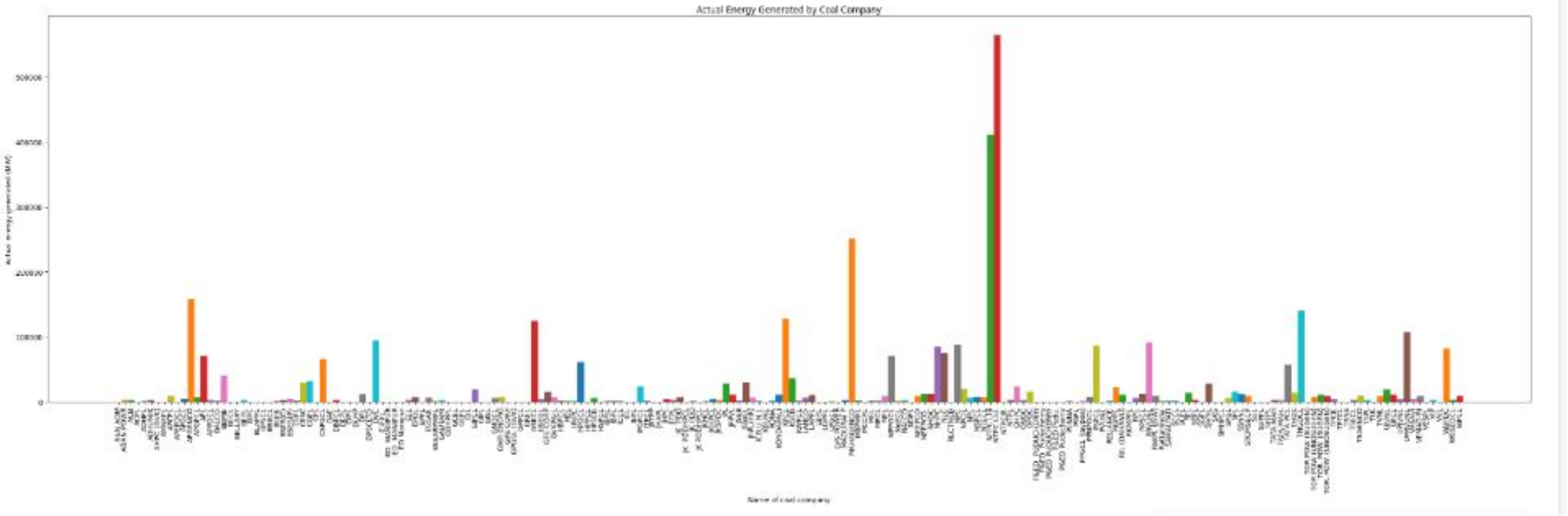
- We will need to look at the columns ‘**Name of coal company’**, **‘Actual energy generated(MW), and the ‘Difference’ column.**
- We group the dataset using the name of the coal company and calculate the total energy generated in MW. I formed two separate plots to do these calculations.
- In one plot, I used all the data. In the other plot, I removed the data of plants operating above capacity
- The difference plot clearly tells us about the amount of energy generated by each coal company that’s technically not possible due to capacity limits.
- There are a total of **204 coal companies.**

Plots



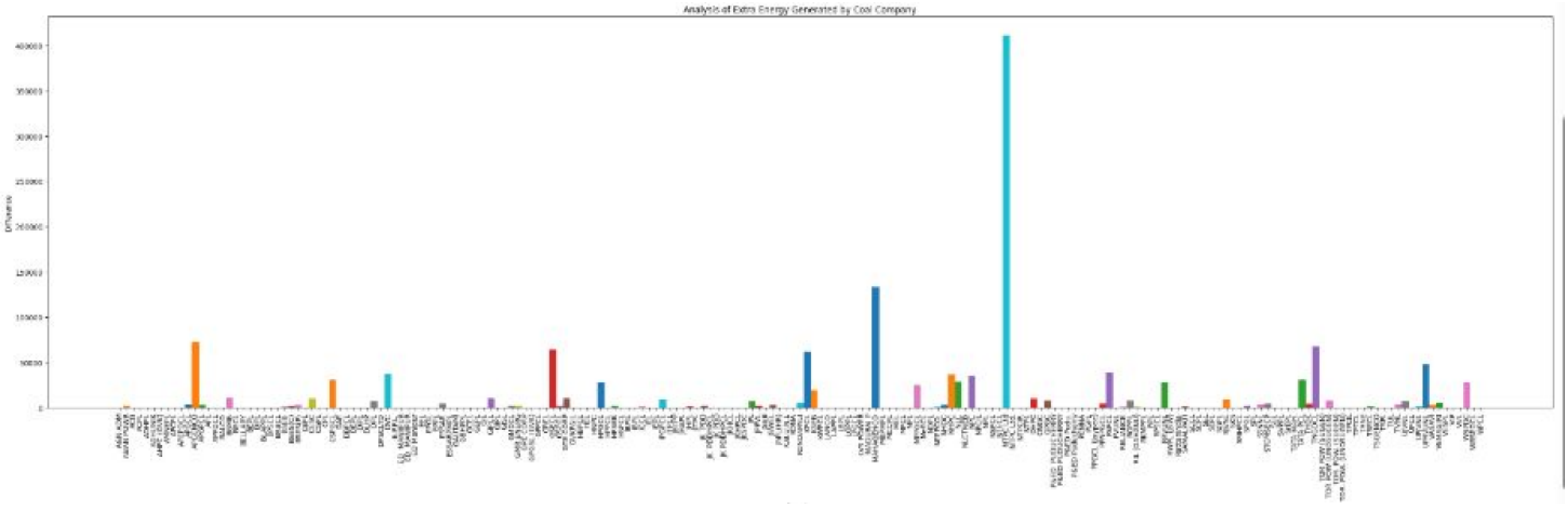
- The above plot shows us the energy generated by all plants operating within capacity limits for each coal company

Plots



- The above plot shows us the energy generated by all plants for each coal company.

Plots



- Each bar represents the difference between the Actual amount of energy generated by each coal company in all plants and the plants operating within capacity limits.
- The higher the bar, the greater the amount of energy produced while operating above plant capacity.

Conclusions

1.

- There are a total of **203** coal companies distributed among different sectors producing energy. The amount of energy generated by each company is distributed in a fairly wide range indicating that the companies have major variation in terms of size, and funding.

2.

- The Difference plots shows that some companies like **NTPC Ltd, APGENCO, and GSECL** among others are producing more 30000 MW each above capacity which is extremely dangerous and could damage the plants. NTPC Ltd is in fact producing **4 lakh MW** above capacity.

3.

- There are **2484 plants** operating under NTPC Ltd operating above capacity of which **2461** are thermal and **23** are Hydro.

Part V

(e) What were the different types of fuels used in the various power plants part of the NDAP Dataset?

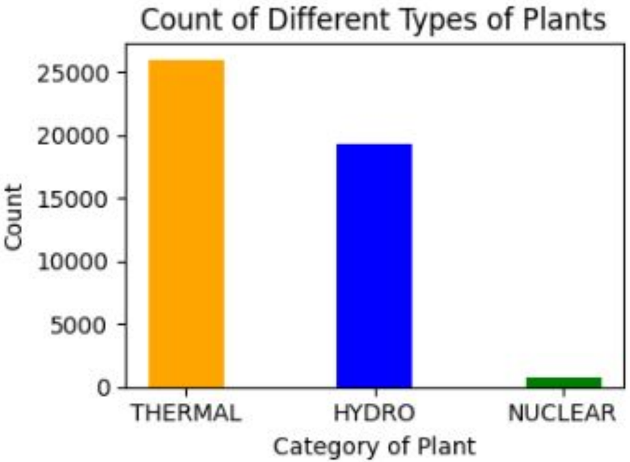
Solution :

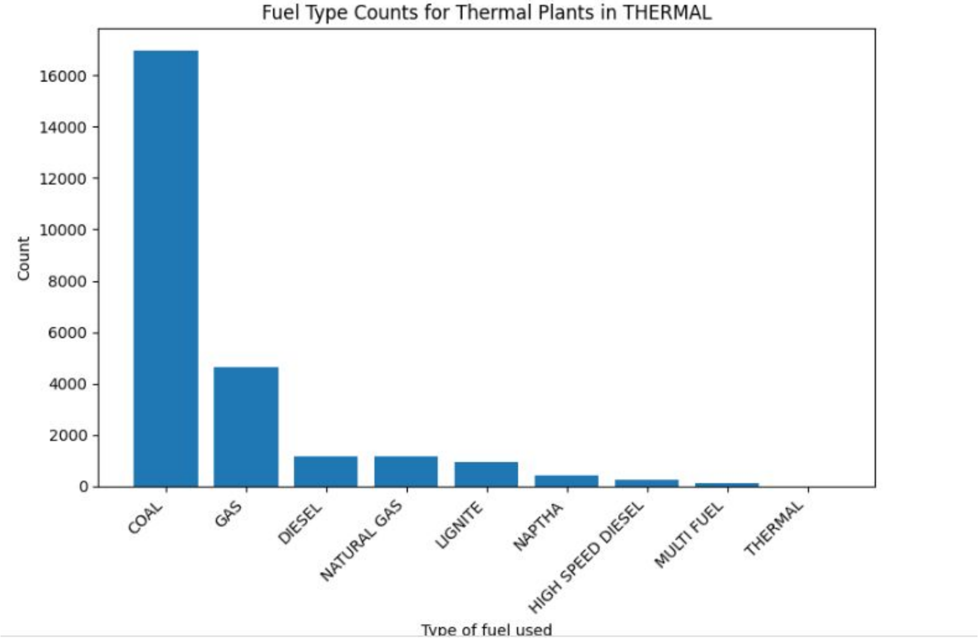
- For the above task we must look at the columns **'Category of Plant'** and **'Type of fuel used'.**
- Upon looking at these 2 columns, we can conclude that there are 3 types of plants :

1. Thermal Power Plants
2. Hydroelectric Power Plants
3. Nuclear Power Plants

- In thermal power plants, there were a variety of fuels used which I also a closer look at.

Plots





- The first plots tells us that most plants in India are either thermal or hydroelectric.
- Nearly **70%** of thermal power plants are powered by coal. Ideally, we need a shift towards renewable sources of energy like biogas or Solar energy. These plots don't consider plants operating above capacity.

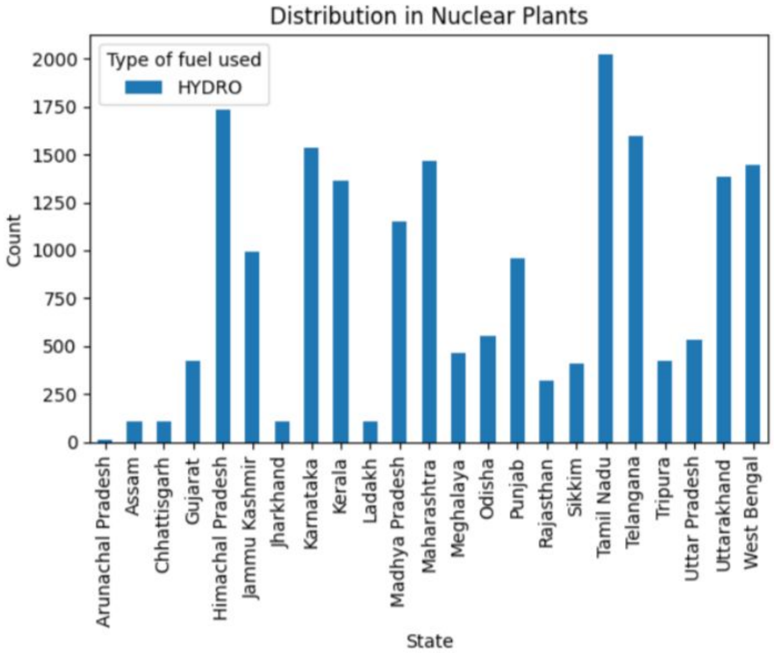
Part VI

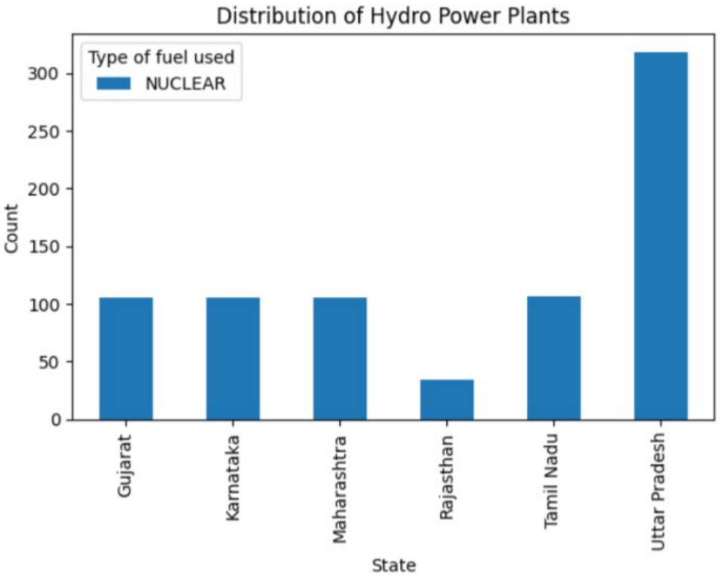
(e) What was the distribution of different types of power plant across the country?

Solution :

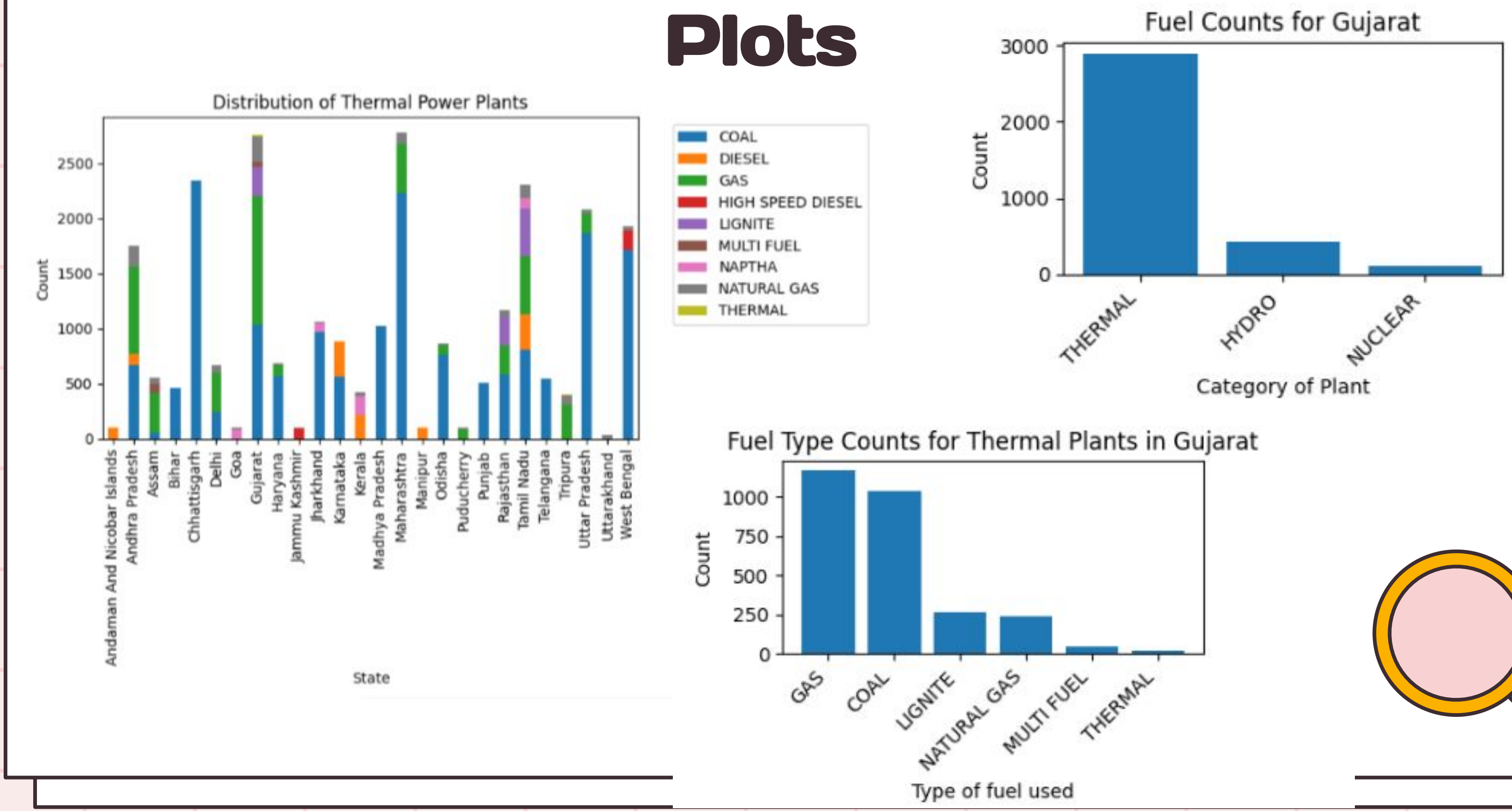
- For the above task we must look at the columns **'Category of Plant'** ,**'Type of fuel used', and 'State.**
- The first step is to group the dataset according to State and use the sum aggregation function to count the number of occurrences of each power plant state-wise.
- We look at the plots generated for the above step and draw some conclusions.
- There are various fuel types used in thermal power plants and thus we dive deeper here and analyse the count of fuel types used state-wise in thermal power plants as well.

Plots


Distribution in Nuclear Plants
Type of fuel used
HYDRO
Count
2000
1750
1500
1250
1000
750
500
250
0
Arunachal Pradesh
Assam
Chhattisgarh
Gujarat
Himachal Pradesh
Jammu Kashmir
Jharkhand
Karnataka
Kerala
Ladakh
Madhya Pradesh
Maharashtra
Meghalaya
Odisha
Punjab
Rajasthan
Sikkim
Tamil Nadu
Telangana
Tripura
Uttar Pradesh
Uttarakhand
West Bengal
State


Distribution of Hydro Power Plants
Type of fuel used
NUCLEAR
Count
300
250
200
150
100
50
0
Gujarat
Karnataka
Maharashtra
Rajasthan
Tamil Nadu
Uttar Pradesh
State

Plots


Distribution of Thermal Power Plants
Count
0
500
1000
1500
2000
2500
State
Andaman And Nicobar Islands
Andhra Pradesh
Assam
Bihar
Chhattisgarh
Delhi
Goa
Gujarat
Haryana
Jammu Kashmir
Jharkhand
Karnataka
Kerala
Madhya Pradesh
Maharashtra
Manipur
Odisha
Puducherry
Punjab
Rajasthan
Tamil Nadu
Telangana
Tripura
Uttar Pradesh
Uttarakhand
West Bengal
COAL
DIESEL
GAS
HIGH SPEED DIESEL
LIGNITE
MULTI FUEL
NAPTHA
NATURAL GAS
THERMAL
Fuel Counts for Gujarat
Count
0
1000
2000
3000
THERMAL
HYDRO
NUCLEAR
Category of Plant
Fuel Type Counts for Thermal Plants in Gujarat
Count
0
250
500
750
1000
GAS
COAL
LIGNITE
NATURAL GAS
MULTI FUEL
THERMAL
Type of fuel used

Conclusions

1.

- Maximum hydroelectric plants exist in Tamil Nadu, Telangana, and Himachal Pradesh. Rivers like Cauveri, and Bhavani must be a great source for these. Telangana is a surprise and it may be that most of these are small plants.

2.

- Some of the States near the Himalayas have very few hydroelectric plants like Arunachal and Assam . Infrastructure and funding issues in these smaller states may be a major factor. There is room for improvement here

3.

- In well-developed states, a variety of fuels are used across power plants while in smaller states generally most thermal plants used Coal or Gas only.

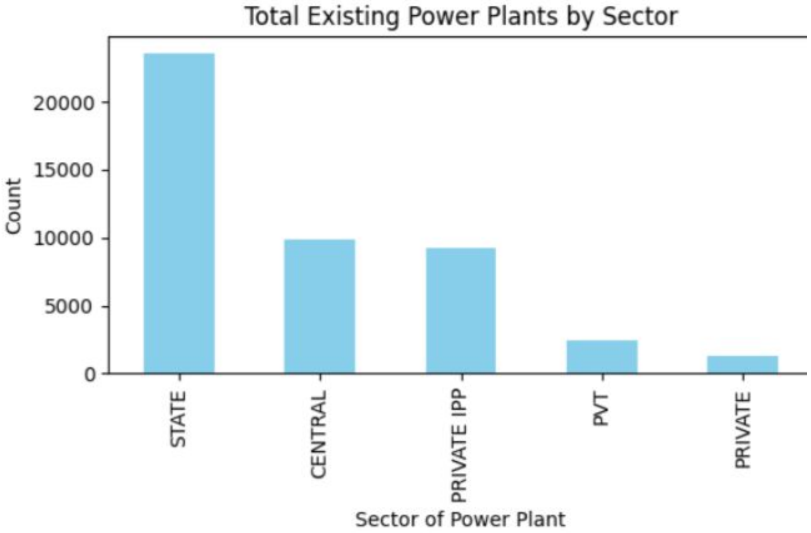
Part VII

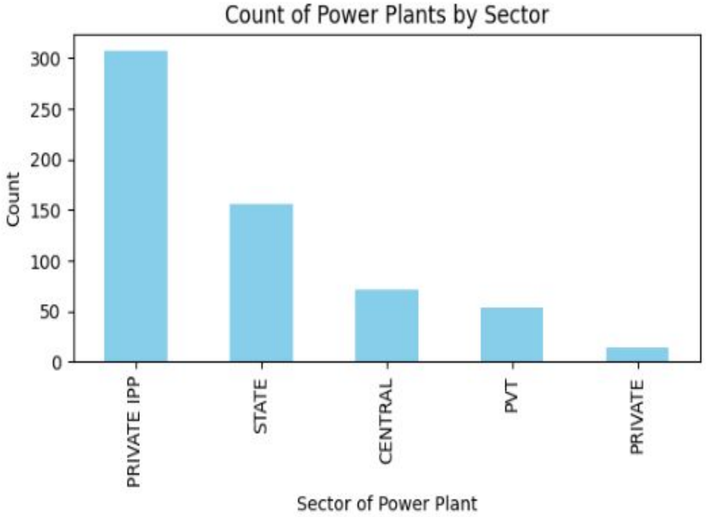
(g) There were scenarios in which plants are operating above there Installed Capacity. What is the distribution of this across sectors?

Solution :

- We form the **Difference** column by taking the difference of values in the column **‘Installed Capacity’ and ‘Actual energy generated(MW)’.**
- For the above task we must look at the columns **‘Difference’,’Sector of Power Plant’, and ‘Actual energy generated(MW)’.**
- We first extract the section of the dataframe in which the value in the **Difference** column in **negative**.
- We then group by sector and count the number of occurrences of each sector in this dataframe fragment.
- We also plotted the energy generated in power plants in which the energy generated was less than or equal to Installed Capacity to get an idea of which plants are generating more energy in a feasible manner.

Plots


Total Existing Power Plants by Sector
Count
20000
15000
10000
5000
0
STATE
CENTRAL
PRIVATE IPP
PVT
PRIVATE
Sector of Power Plant


Count of Power Plants by Sector
Count
300
250
200
150
100
50
0
PRIVATE IPP
STATE
CENTRAL
PVT
PRIVATE
Sector of Power Plant

Conclusions

1.

- Out of the **603** plants operating above capacity, nearly **50%** are Private IPP(Independent Power Producers). This indicates regulations need to imposed more strictly in the private sector to ensure safe operation of the plants. An independent regulatory could be set up for this purpose

2.

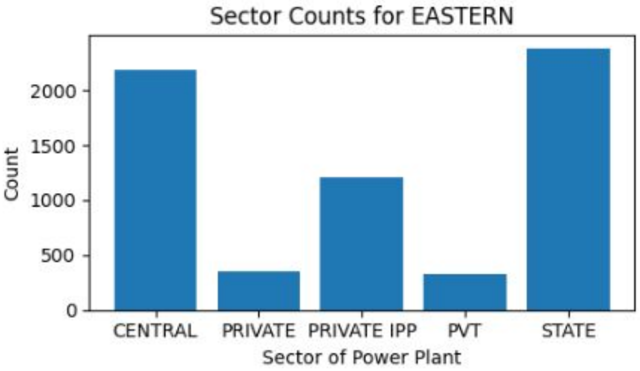
- Maximum amount of energy is produced in a feasible manner in Central and State Plants indicating efficiency in the operations of the CEA

Part VIII

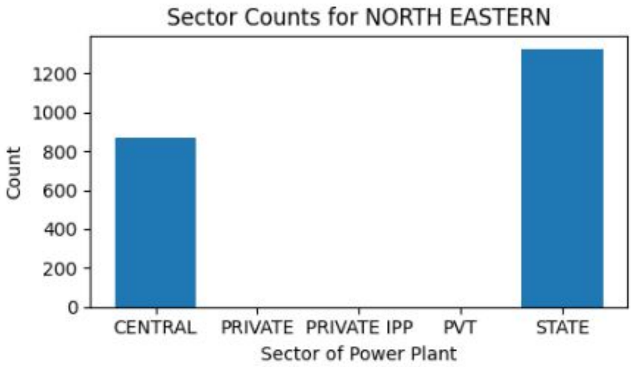
(h) What is the distribution of Power Plant Sectors across the states and regions in the country?

Solution :

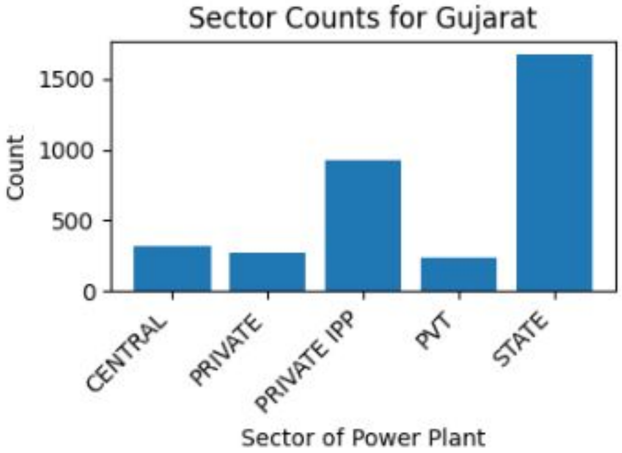
- We will need the columns ‘**Sector of Power Plant**’, ‘**State**’, ‘**Region**’.
- This is a fairly simple task which gives us an idea of the sector-wise distribution of plants across the country.
- We will group by State and Region to form 2 separate groupby objects and count number of occurrences of each sector in every state iteratively.
- We then generate one plot for each state which tells us how many power plants exist in each sector of the state.
- We do the same thing for Region as well and plot it too.



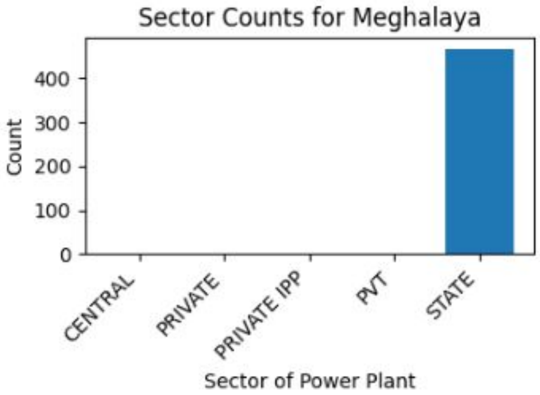
Sector Counts for EASTERN



Sector Counts for NORTH EASTERN



Sector Counts for Gujarat

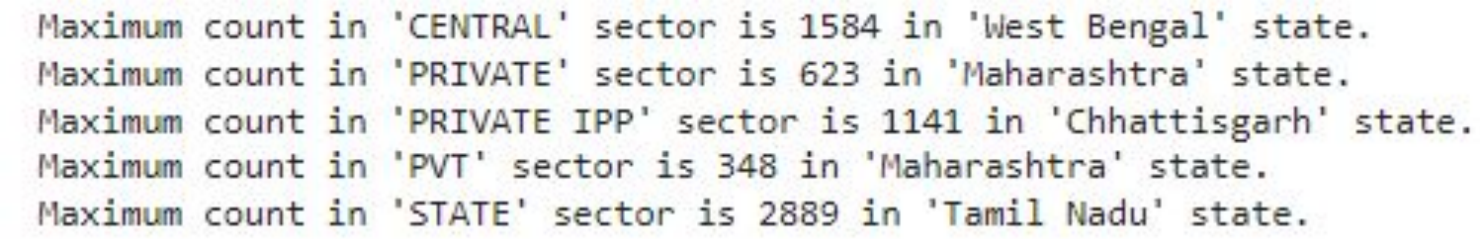


Sector Counts for Meghalaya

Conclusions

1.

- The Central and State owned plants are well distributed across the country. However, the private sector is mainly concentrated on the Western and Eastern Sides of the country with only few in the North and South. There are absolutely none in the North East. This could be to lack of infrastructure and funding in these states.Further sources should be consulted to understand why this is so.



```
Maximum count in 'CENTRAL' sector is 1584 in 'West Bengal' state.
Maximum count in 'PRIVATE' sector is 623 in 'Maharashtra' state.
Maximum count in 'PRIVATE IPP' sector is 1141 in 'Chhattisgarh' state.
Maximum count in 'PVT' sector is 348 in 'Maharashtra' state.
Maximum count in 'STATE' sector is 2889 in 'Tamil Nadu' state.
```

- I also calculated the States in which each sector had maximum number of plants.

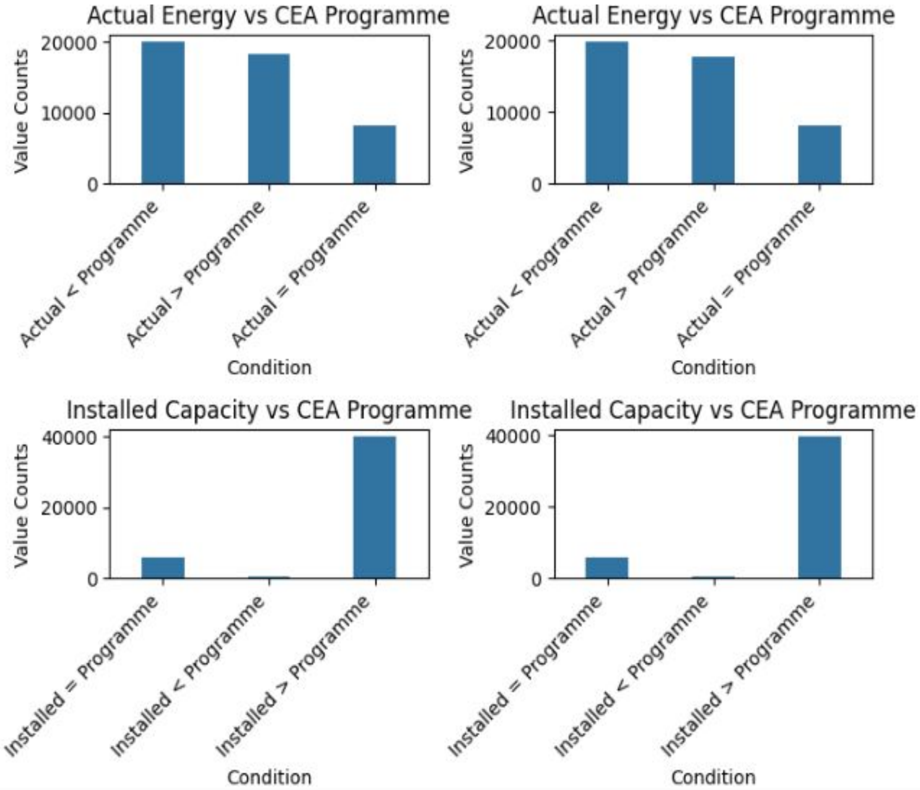
Part IX

(i) Upto what extent are the CEA Generation Programme Regulations being maintained?

Solution :

- We will need the column **‘Actual energy generated(MW)’**, **‘Installed Capacity’**, and **‘Generation Programme is prepared by CEA’.**
- We first compared the Installed Capacity and the CEA Generation Column and split the dataframe into 3 parts based on which value was larger in a particular power plant.
- We then repeated the same procedure to compare **Actual energy generated(MW) and CEA Generation** column.
- We then plotted the entire dataset once, and separately repeated plots after removing plants which were operating above installed capacity. The distribution was almost the same as the fraction removed was very small.

Plots



Actual Energy vs CEA Programme

Actual Energy vs CEA Programme

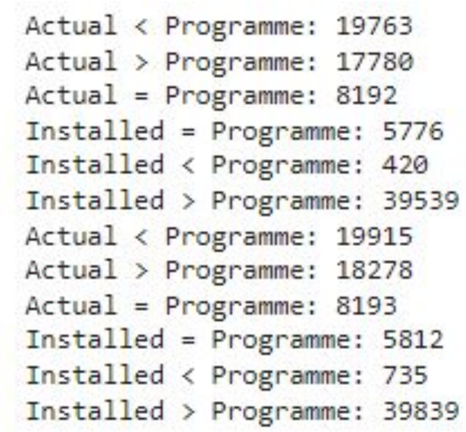
Installed Capacity vs CEA Programme

Installed Capacity vs CEA Programme

Conclusions

1.

- A vast majority of plants are producing energy at a rate much lesser than that specified by the CEA in spite of having sufficient capacity to do so. There are a small fraction of plants that need to undergo renovation as well to meet the programme requirements



```
Actual < Programme: 19763
Actual > Programme: 17780
Actual = Programme: 8192
Installed = Programme: 5776
Installed < Programme: 420
Installed > Programme: 39539
Actual < Programme: 19915
Actual > Programme: 18278
Actual = Programme: 8193
Installed = Programme: 5812
Installed < Programme: 735
Installed > Programme: 39839
```

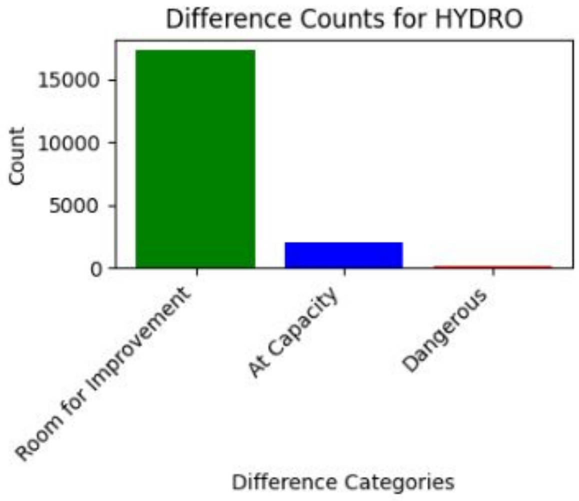
Part X

(i) Is there a particular type of plant(thermal, nuclear,hydro) particularly operating above capacity?

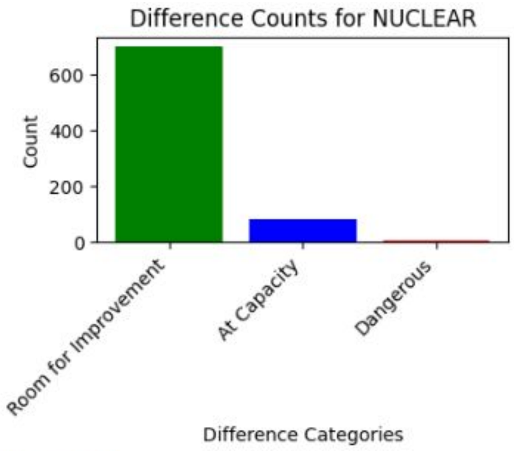
Solution :

- .Just as we did previously to analyze this distribution with respect to location, we form the Difference column and group the dataset using Category of Plant and the Differences. We count the number of occurrences of difference being positive, negative, and zero and plot these occurrences for each type of plant.

Plots

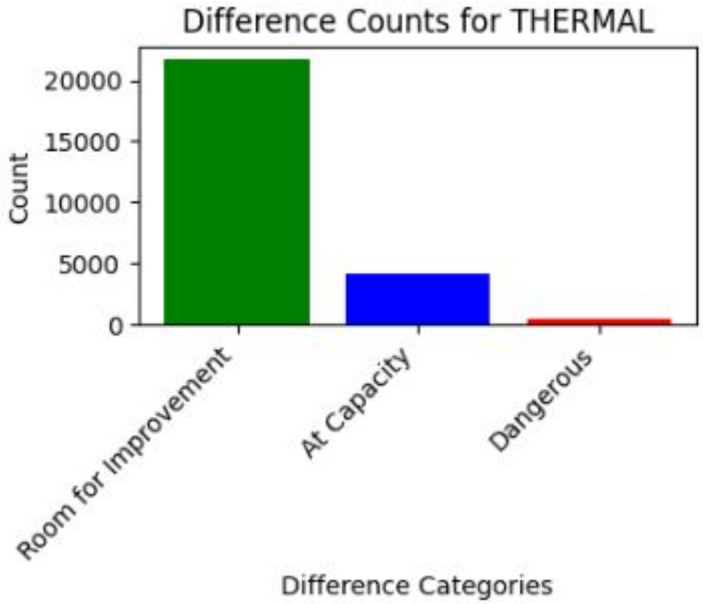


```
[17246, 1967, 116]
```



```
[697, 80, 6]
```

Plots



```
[21644, 4149, 481]
```

Conclusions

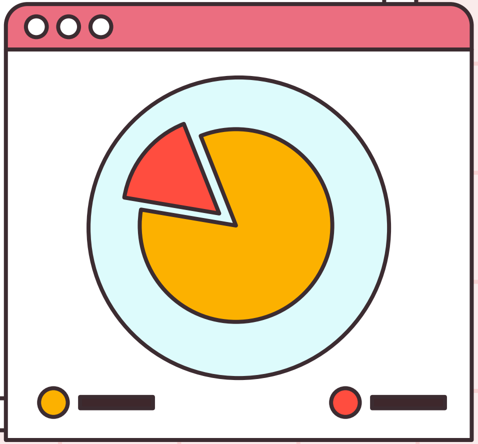
1.

- It is majorly thermal power plants that are operating above installed capacity. It is definitely a relief that there are only 6 nuclear plants doing so. They immediately must be checked up upon as nuclear disasters are quite dangerous.
- After some further analysis, I have also found out all the data corresponding to these plants is from April 2008 and that these plants are located in UP(3) Gujarat(1), Karnataka(1), and Maharashtra(1).

02

Pollution Trends

Correlating Power Plants with Pollution:
Unveiling the Hidden Trends!



Problem Statement

- We have previously analyzed the NDAP Dataset in great detail looking at both regional variation across the country as well as temporal variation among other things.
- With the consumption of fossil fuels in these plants, there's bound to be a significant amount of pollution.
- We must keep a check on these pollution levels and analyze them alongside plant data to understand how we must operate in order to remain safe

Addition Data Requirements for this would be :

1. Historical air quality data (PM2.5, PM10, NOx, SO2 levels).
2. Emission data specific to power plants (CO2, SOx, NOx emissions).
3. Data regarding Vehicular Emissions in areas around power plants

SOLUTION

Sources of this new data would include :

- Central Pollution Control Board (CPCB): Air quality data, emissions data.
- State Pollution Control Boards: Regional pollution data.
- World Health Organization (WHO): Global pollution data and trends.
- NGOs and Research Institutes: Studies and reports on pollution trends in India.

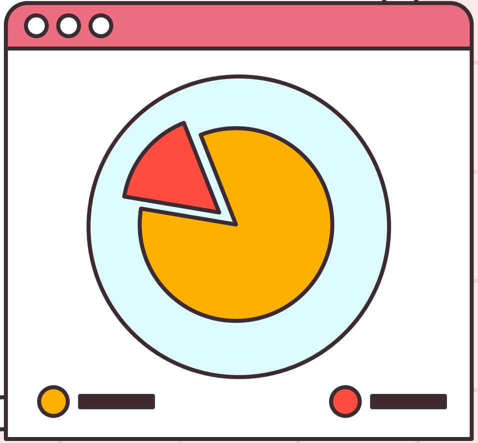
To Carry out this Correlation, I would use the following methodology :

1. Form a database of the pollution data and create a dataframe using it. Then, merge this dataframe with the NDAP Dataset along a common column such as Year or State.
2. Calculate correlations between variables such as energy generation, type of fuel used, and pollutant levels.
3. Visualize and analyze spatial patterns of pollutant concentrations in relation to the locations of power plants.
4. Analyze temporal patterns in energy generation and emissions from power plants over different time scales (yearly, monthly, seasonal).
5. Analyze pollution levels by company, and sector as well.

03

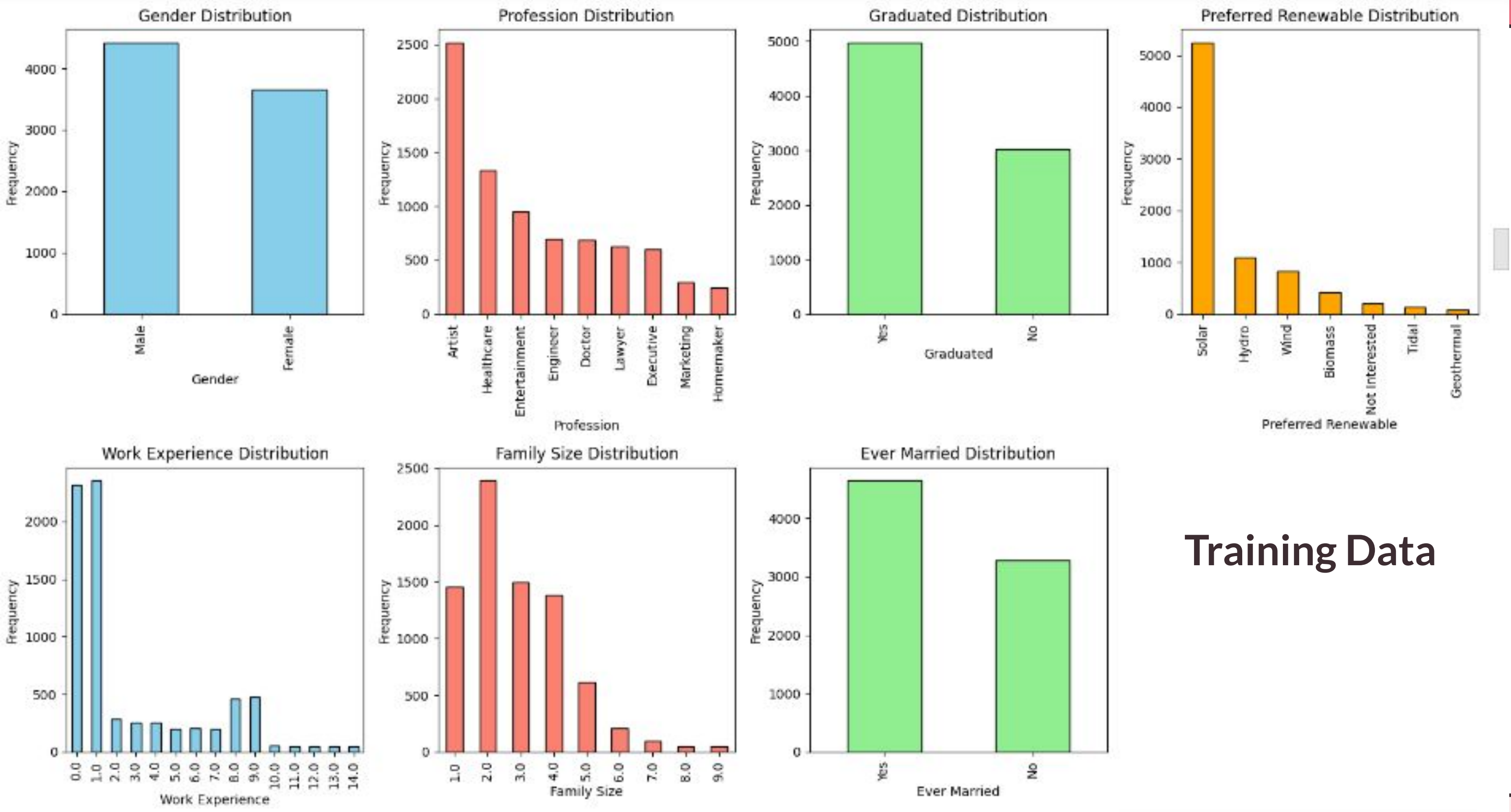
Customer Segmentation

Predicting Customer Segments for Optimal Marketing Strategies

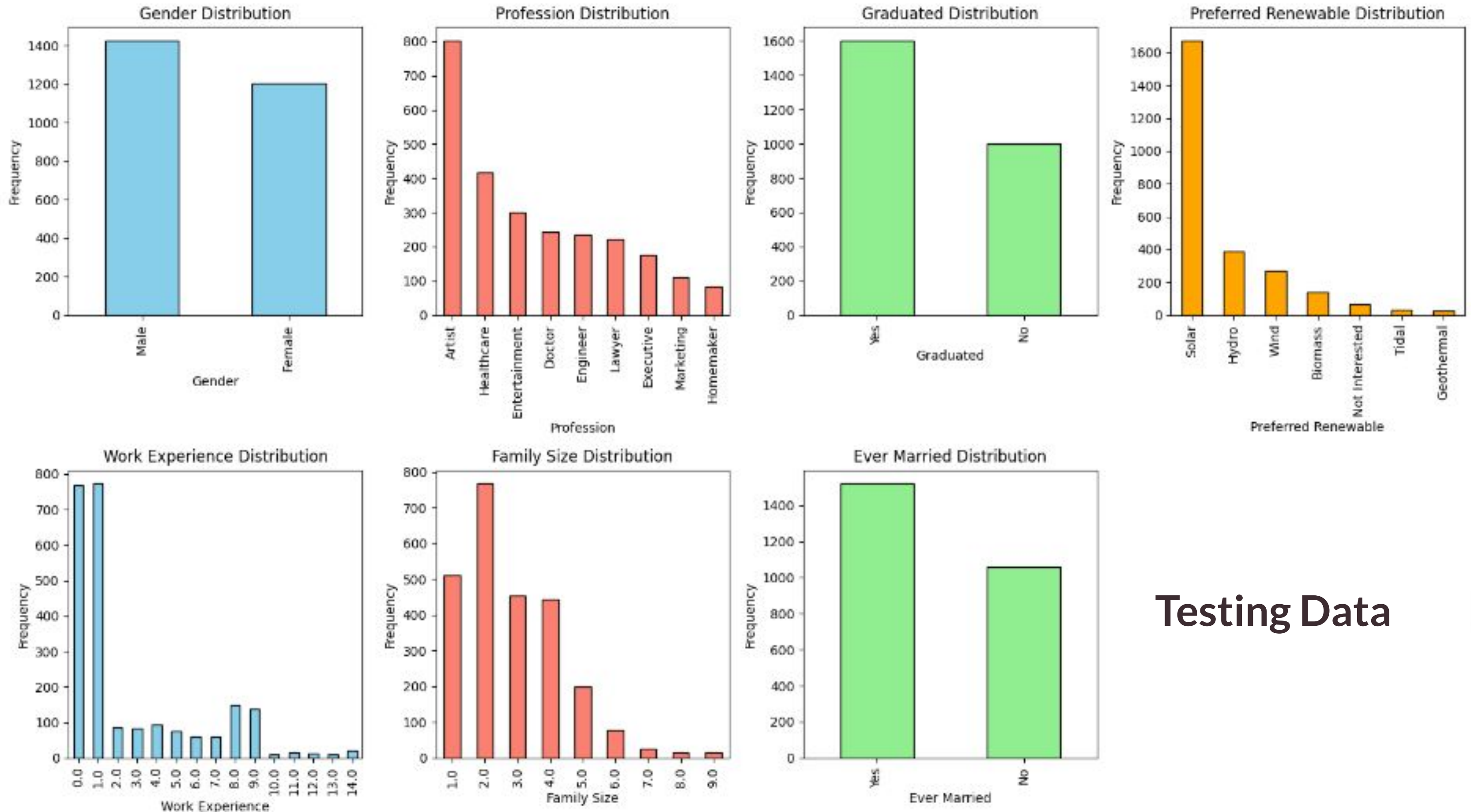


Analysis

- The given problem statement requires us to classify customers into four segments (A, B, C, D).
- This segmentation enables the formation of tailored marketing strategies that specifically appeal to each group.
- The dataset consists of customer information across various demographic and behavioral attributes, The dataset consisted of 8068 rows and 9 features + the final grouping column of the customers.
- One point to note was the lack of correlation between the variables in the dataset. For example, a 56 year old may have had zero work experience and a 50 year old may have had 13 years of work experience. These kind of relations between columns was highly prevalent making imputation of one column based on others was a poor idea.
- I could either fill the null values in the dataset with the mode or drop them entirely.
- I decided to do the latter following which I scaled the data in numerical columns
- I have attached a screenshot of the plots for each of the features which helps me justify my choice of using imputation with mode



Training Data



Testing Data

SOLUTION

- In Conclusion, I decided the best action would be to fill NaN values in each column with Mode given the highly skewed nature of this dataset and also the test data. Mean is not suitable due to the skewed nature of the dataset.

- The next step was encoding the categorical columns in the dataset. I made use of LabelEncoder from scikit-learn to do so. I also scaled the numerical data using StandardScaler.

- The dataset was now ready so the next step was to split the data into the training and testing set following which the models would be trained.

- I made use of GridSearchCV to test multiple supervised learning models like RandomForest, XGBoost, and DecisonTree.

- The accuracies obtained were of similar orders with XGBoost giving the best performance. Thus, I used XGBoost to train the model and then used this model to predict groups for the test data. The test set accuracy was **53.0%.**

SOLUTION

- To summarize, I had 2 options :

1. Drop the rows with NaN values, and scale the data before creating and training my model.

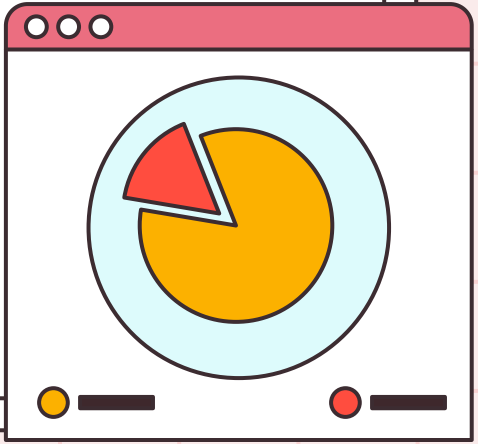
2. Fill the numeric columns which have NaN values with a SimpleImputer using a Mode strategy . Similarly, the non-numeric columns also have some NaN values which can be filled with the mode or the most frequent item in that column.

- The values in the numeric columns were highly skewed towards one end which is why I did not consider Mean Strategy in my SimpleImputer
- Given this skewed nature of the dataset and the test dataset, filling NaN values with Mode seemed like a good option.
- Also dropping 1400 columns is a tremendous loss of data so i felt the trade off of making the model skewed while keeping more data was a good one.
- **Assuming the entire market in India follows this trend this model will work well on any test data.**

04

Unsupervised ML

Unlock customer segmentation mysteries with clustering



SOLUTION

- The given problem statement required us to classify the test data in the absence of the training set which was used in Q3.
- I would make use of a **clustering algorithm** such a situation.
- Clustering helps uncover patterns and similarities among customers without prior knowledge of their groups.

- However, before implementation of the clustering algorithm, I would first carry out the following steps :

1. **Feature Selection:** Identify relevant features that describe customers' behaviors, demographics, purchase history, etc. These features should be chosen based on their potential to differentiate or segment customers. In this case I will directly use the test dataset.

2. **Data Preprocessing**: Clean and preprocess the data to handle missing values, scale numerical features if necessary, and encode categorical variables.

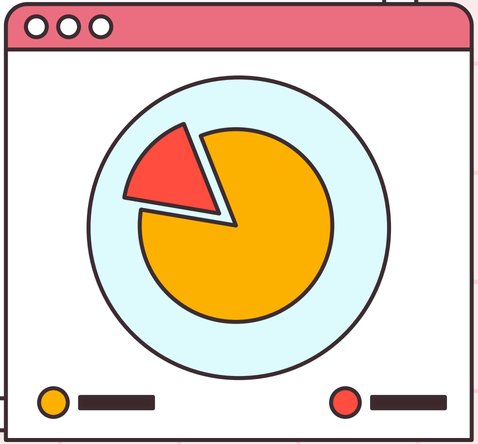
SOLUTION (cont.)

- I would choose one among multiple clustering algorithms like K-means Clustering, Hierarchial Clustering, and DBSCAN.
- Once the clusters are formed, I would analyze them to understand the characteristics and behaviors of customers within each cluster. I'll look for distinctive traits or patterns that differentiate one cluster from another.
- The next step would be to assign meaningful labels to the clusters to identify the customer segments.
- **I have used K-means clustering to implement in Q5.**

05

Customer Segmentation

Clustering the data!!

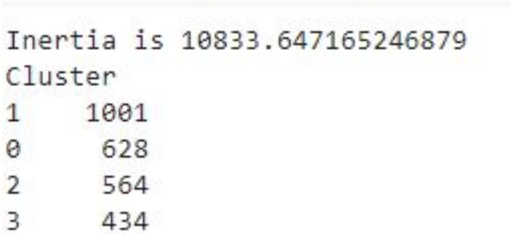


Code Implementation

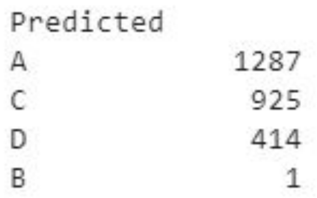
- Now, I have explained how I would go about forming groups in case I do not have any training data.
- I have now implemented that in the form of code using K-Means Clustering.
- Initially, I have carried out the same data preprocessing I did in Q3 to get the dataset ready.
- I used SimpleImputer to replace NaN occurrences in each column with the mode following which i encoded the categorical columns and scaled the numeric columns.
- After this I used Kmeans clustering to create 4 clusters.

Comparison with Q3

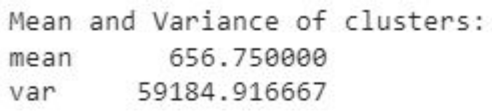
- Once I finished the clustering into groups, I compared the results obtained with that of Q3.
-



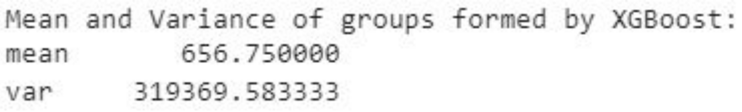
```
Inertia is 10833.647165246879
Cluster
1    1001
0     628
2     564
3     434
```



```
Predicted
A            1287
C             925
D             414
B               1
```



```
Mean and Variance of clusters:
mean       656.750000
var      59184.916667
```



```
Mean and Variance of groups formed by XGBoost:
mean       656.750000
var     319369.583333
```

Comparison with Q3

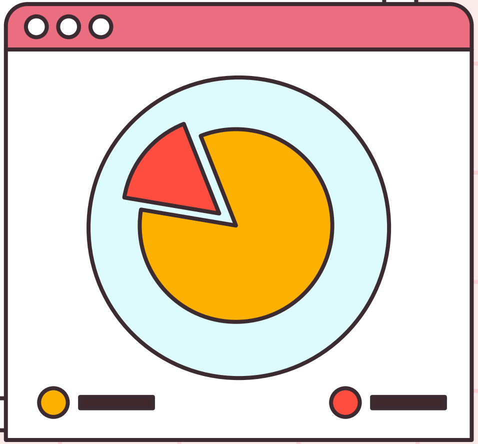
- An Interesting observation is that the mean is identical in both cases. This could be due to to the fact that the dataset is exhibiting some clear patterns which is causing clusters and the Decision Boundary to separate the dataset in a somewhat similar manner

- There is a large difference in the variation which could be due to irregular shape of some clusters which causes KMeans to struggle to form accurate centroids

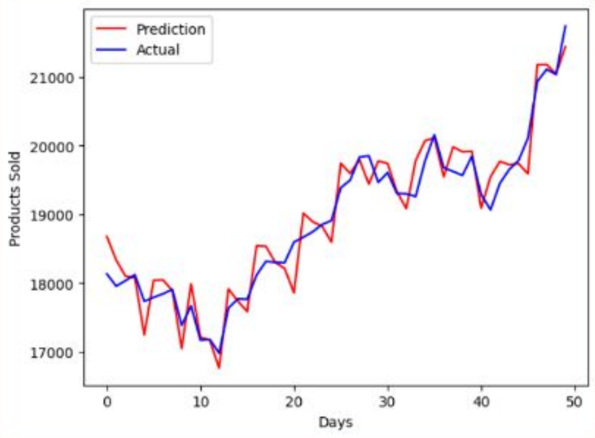
06

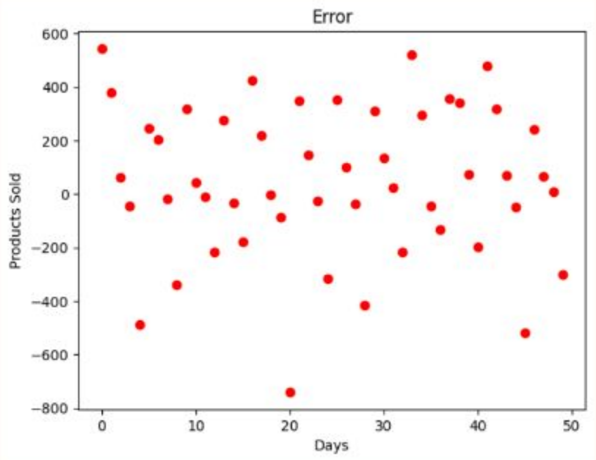
Model Validation

What’s wrong with the Model????



Problem Statement





- A model has been created to predict and analyze daily production based on factors like weather, season, and other variables.. However, it is performly poorly as seen above. Why is that so?

Analysis

- Potential issues with the model include :

1. **Feature Selection :** Though this model follows the general trend of actual production, but there are significant minor deviations at regular intervals. There may be high correlation among features . When two features are similar, they provide the same information to the model. Thus, the model may be overemphasizing some particular feature
2. **Overfitting (High Variance) :** The model we are using may be too complex and capturing noise in the data along with the underlying patterns. This is evident from the random distribution of the errors and the deviations of predicted values from true values
3. **Underfitting (High Bias) :** The model we are using may be too simple and unable to capture the underlying patterns in the data.
4. **Poor Hyperparameter Selection :** It is possible that the selection of hyperparameters for the model is quite poor as well due to which the model is performing poorly.
5. **Model Selection and Complexity :**The model being used may not be the right one here. We should try more models and analyze the results.

Solution

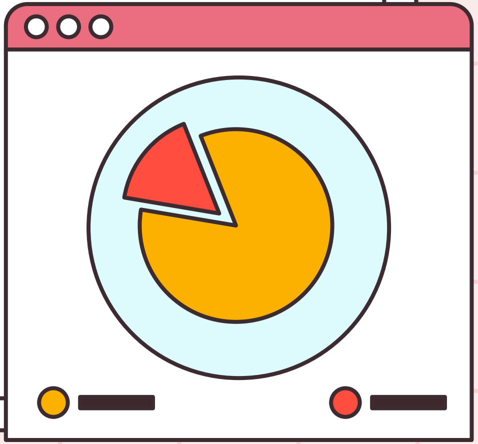
- Possible ways to solve these issues include :

1. **Dimensionality Reduction and Feature Engineering :** I would check the correlation between features in the dataset and drop redundant features. I would try to analyze the dataset further as well to see if any important features are missing. With some domain knowledge of the data, I will decide on new features, if any.
2. **Regularization of the Data :** I would try to use Regularization techniques to help reduce overfitting in the data. Also, if it was a neural network, Adding **EarlyStopping,** and **Dropout** layers should also help improve the model.
3. **Hyperparameter Tuning** : I would use **GridSearchCV** to test out various different models with different parameters in order to identify the best possible model and parameters.

07

Spam Classifier

Time to clean up your inbox!



SOLUTION

Data Preprocessing

- We are provided with a training set of 4000 emails which are to be used to predict whether 1000 emails in the test set are Spam or Not Spam.

- The first step is to do some preprocessing to get the dataset ready for training.

- Firstly, I encoded the Target Variable to show 1 for Spam 0 for Not Spam

- The next step is to clean the emails to make the understandable.

SOLUTION (cont.)

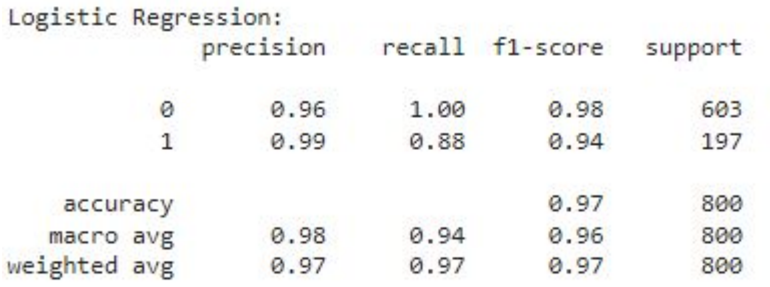
Data Preprocessing

- I have made use of basic NLP tools from the nltk library to clean the messages.
- I removed all special characters, and converted all messages to lowercase.
- I also removed stop words, which would be common to both spam and non-spam mails which will allow better classification.

SOLUTION (cont.)

The Model

- I analyzed the performance of different models like Naive Bayes, Logistic Regression, Decision Tree, and RandomForest on the dataset.
- I looked at the parameters Precision, Recall,, F1-Score, and Accuracy to come to a decision.



```
Logistic Regression:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98       603
           1       0.99      0.88      0.94       197

    accuracy                           0.97       800
   macro avg       0.98      0.94      0.96       800
weighted avg       0.97      0.97      0.97       800
```

SOLUTION (cont.)

The Model

- Among all the best model was Logistic Regression. It had **99%** accuracy in identifying correct cases. It gave minimum false positives.
- It had an F1_Score of **0.94** indicating good balance between precision and recall.
- It also had the highest accuracy among models on the training data of **97%**

THANK YOU