Data analytics and visualization team Assignment

Question 1. Ask 10 reasonably involved questions and try to answer them by analysing the Dataset.

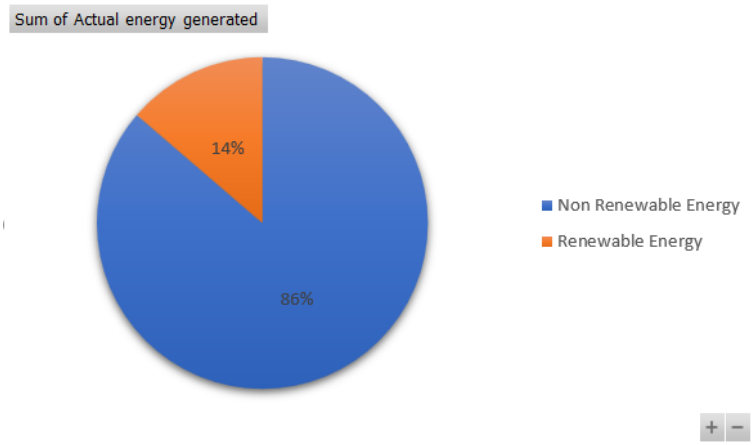
Ans.

1. what is the current total capacity of renewable energy and amount of power generation generated?

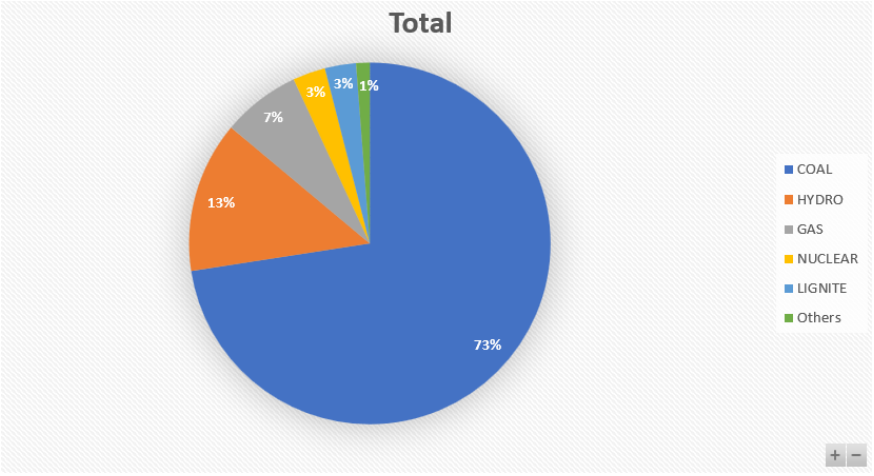
   As of 2017:
   sum of Actual energy produced = 21098.97 GWH
   Sum of installed capacity = 132431.26 MWH

   While the total installed capacity of renewable energy so far
   = 4329804.62 MWH

2. What percentage of India's total power generated comes from renewable energy sources?



3. What is the distribution of power generation in India based on the type of fuel used?



4.
5.

4. What is the dominant type of fuel used among renewable source of energy?

   Sum of Actual energy generated for Hydro: 1819268.22 GWH

   Sum of Actual energy generated for NAPTHA: 25012.51 GWH

   Hence hydro is more dominant type fuel used among renewable source of energy.

5. Which region/state is having highest energy generation in last 5 years?

Row Labels	Sum of Actual energy generated (GWH)
NORTHERN	280157.29
SOUTHERN	107917.17
WESTERN	71922.13
EASTERN	47994.28
NORTH EASTERN	16821.85

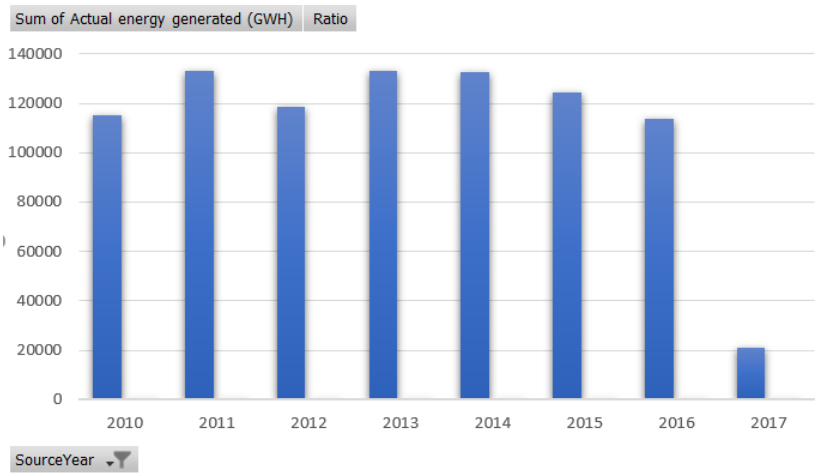
Row Labels	Sum of Actual energy generated (GWH)
Himachal Pradesh	81122.52
Punjab	63061.78
Jammu Kashmir	57694.04
Karnataka	42638.8
Uttarakhand	39661.84
West Bengal	35255.66
Madhya Pradesh	29162.12
Kerala	28116.47

6. Which state comparatively require more capacity?

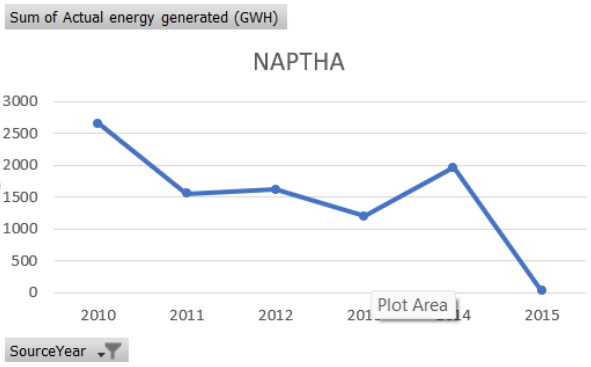
Row Labels	Sum of Actual energy generated (GWH)	Sum of Installed Capacity (MWH)	Ratio
Jammu Kashmir	57694.04	135929	0.4244
Sikkim	14241.38	37387	0.3809
Himachal Pradesh	81122.52	243475.85	0.3332
Assam	1632.74	5000	0.3265
Punjab	63061.78	196481.9	0.321

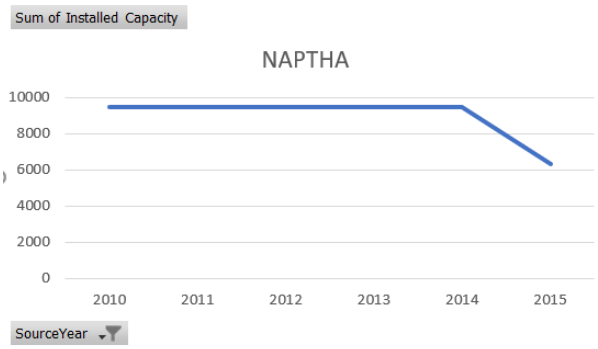
Clearly ratio of energy generated to installed capacity is highest in Jammu Kashmir in recent years, hence it the most likely state to **require more capacity**.

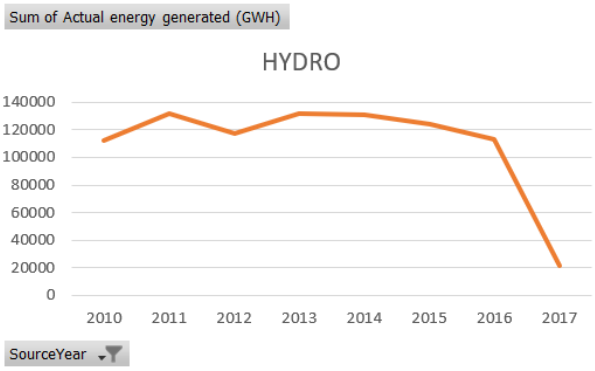
7. What is the trend of India's power generation usage in recent years?

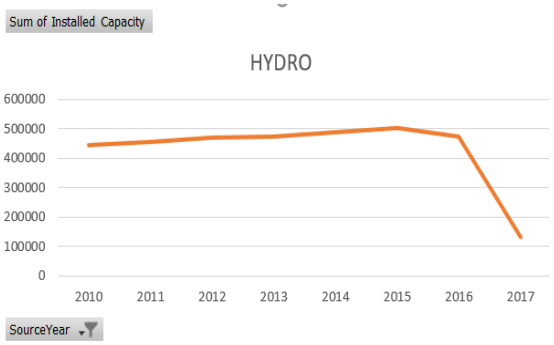


8. Does the dataset indicate any increases in installed capacity/ power generation for specific renewable energy sources?



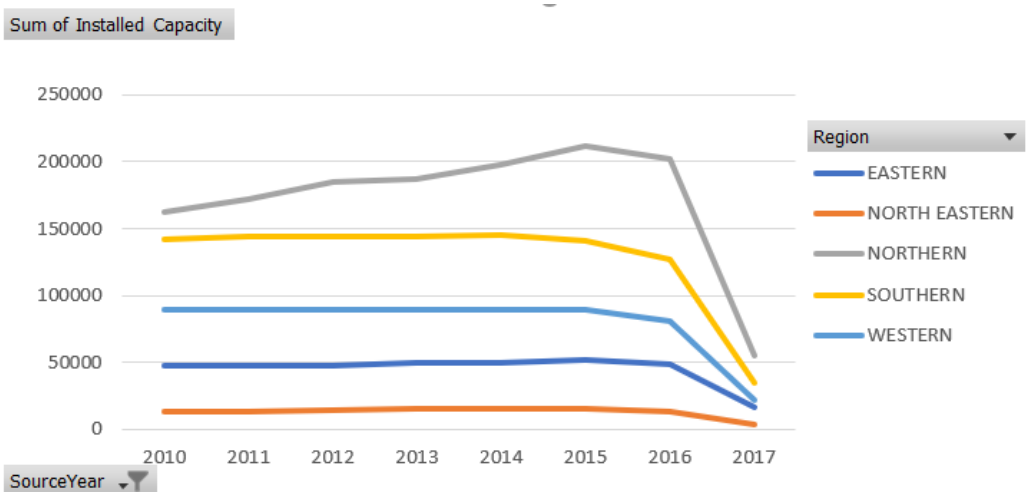






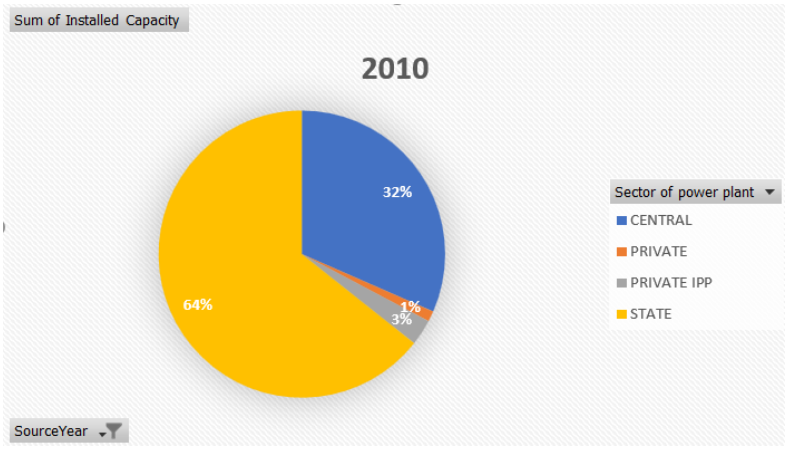
No such increase in installed capacity / power generation for any renewable energy sources

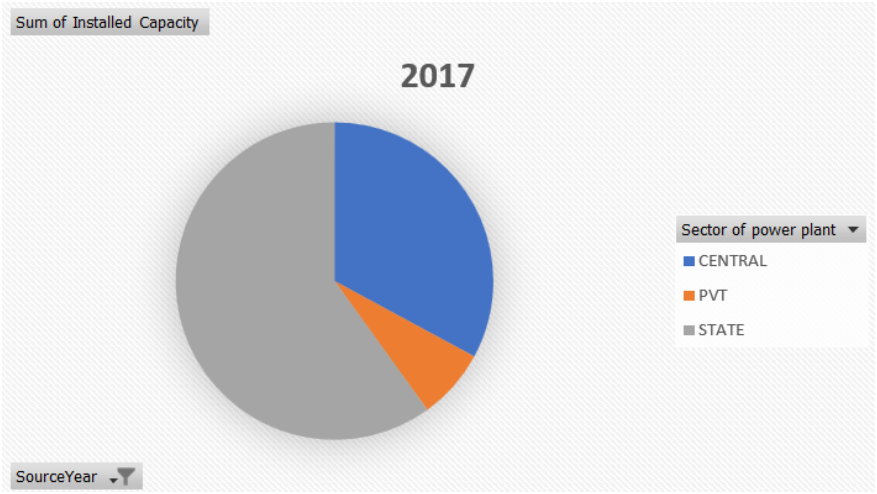
9. Does the dataset indicate any increases in installed capacity for any specific state/ region?



No such increase installed capacity for any specific state/ region.

10. What is the proportion of sectors of renewable sources of energy in last decade?





Question 2. Seeing high levels of pollution in the country, you get curious about the trends in pollution and try to correlate it with the dataset for power plants you are given. How would you go about doing that? Is the current dataset sufficient to identify pollution trends? If not, what additional data would you need, and where would you obtain it? List the potential sources for acquiring this necessary data..

Ans. The current dataset on power plants can somewhat provide some information, but it doesn't directly show us pollution trends but what I got insight from dataset are:

- Filtering the data for power plants that uses polluting fuel types such as coal, gas, lignite, and diesel we can analyse trends in Installed Capacity and Actual Energy Generated for these plants. If we see an increase in capacity or generation, it might suggest a indication with rising pollution levels.
- Looking for power plants that are located near areas with high pollution levels which could indicate that these plants are contributing to the pollution in those areas.

No, the current dataset is not sufficient to identify pollution trends.

We would require dataset of amount of emission of harmful gases and chemical from these plants. Air Quality Data

Source:

1. www.statista.com
2. www.kaggle.com
3. cea.nic.in

Question 3. Implement methods to assist your team in identifying and predicting the appropriate group (A, B, C, or D) for each customer in the test data.

Ans. I will use random forest classifier for classifying dataset into 4 groups i.e. A B C D.
file name: groups_classifier.ipynb

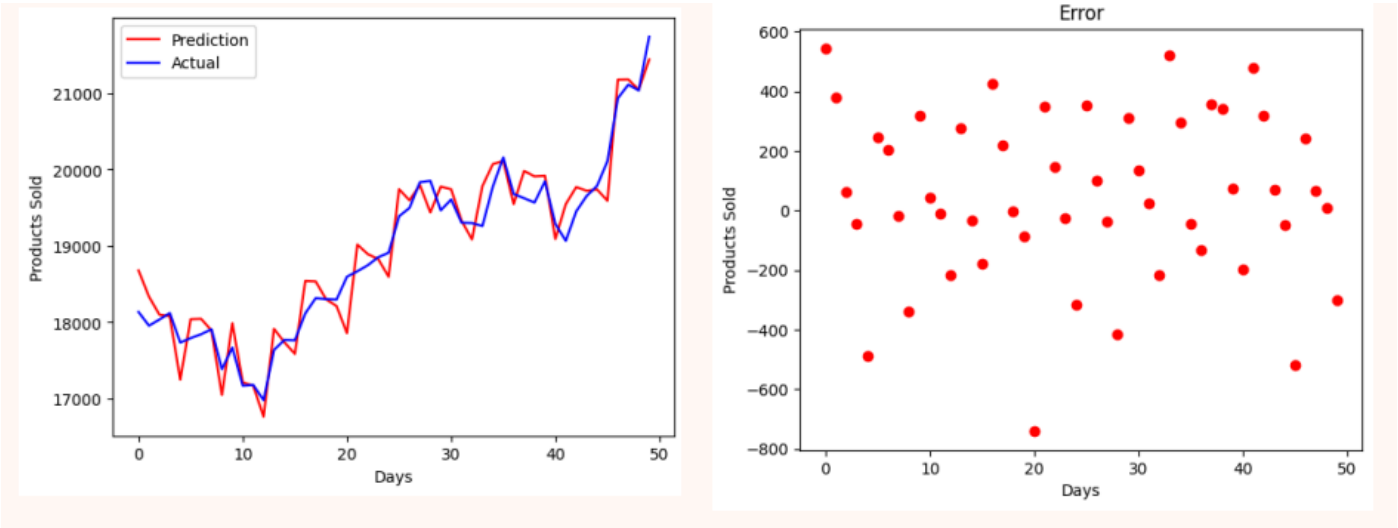
Question 4. In the above case your team already had data from which customers were classified in groups. How will you predict the classes if the groups and it's data weren't available? State the method you would have used.

Ans: In such case, we would approach the problem as a typical unsupervised learning task, specifically clustering. The method I'll be using will be K means clustering.

Question 5. Implement the method described in Q4 and compare the groups formed with A,B,C,D (Q3).

Ans. File name: consumer_clustring.ipynb

Question 6. What do you think might be causing the poor performance of the model? To improve the model' s accuracy, what steps would you take? Provide a detailed justification for each of your proposed methods



Ans. According to me, following can be the reason for poor performance of the model:

- The model might be missing important features that affects production such as market trends, promotional events, holidays, competitor actions. There can be other factors too that maybe influencing the sales internally.
- The model can be underfitted. This could be because the model is too simple or not capturing the complexity of the relationship between the features and sales.
- Poor hyperparameter tuning.

These improvements can be made improve accuracy:

- Analysing the existing data to identify features that correlate with production levels.
- Exploring different model architectures, like for example if it is a simple model, we can switch to a more complex model like a Long Short-Term Memory (LSTM) network.
- We can experiment with different hyperparameter values for the chosen model architecture. As far as I know Grid Search or Randomized Search could be a better option.

BONUS. Develop a model to predict whether each email is spam or not, and use it to classify the uncategorized emails.

Ans. File name: spam_classifier.ipynb