

# BIG DATA ANALYSIS USING IBM CLOUD

## Project's Objective:

The objective of this project is to clearly understand the data analytics concept and how to use the IBM cloud for find the hidden trends and patterns inside the dataset and IBM Watson Studio for visualization like creating graphs and charts. This project also includes how to use the python libraries like numpy (numerical python), pandas (framing the data), matplotlib (plotting the graph) for analyze the data.

## Design Thinking:

Design thinking is the iterative process of how to perform the project.

**Data Selection:** Identify the datasets to be analyzed, such as climate data or social media trends data but our team choose the climate data of analyze.

**Database Setup:** Using the feature code in the skillup course like Introduction to cloud to freely open our IBM cloud and use the IBM db2 and IBM Watson Studio for analyze, store the dataset and visualization the dataset respectively.

**Data Exploration:** In IBM db2 develop queries and scripts to explore the datasets,extract relevant information and identify patterns.

**Analysis Techniques:** Applying the appropriate machine learning techniques to find the hidden patterns in the dataset. Our team use the K-Means Clustering algorithm to find the hidden pattern in the dataset.

**Visualization:** Using the IBM Watson Studio to visualize the dataset like graphs and charts to understand about the hidden trends and patterns in the dataset.

## How our team performs in the Project.

**Step 1-** At first, clearly understand the concepts in each phase.

**Step 2-** After understanding, discuss about the concepts with our team members to perform it.

**Step 5-** Again, review the whole project and ask my mentor to review about our project and corrects mistake if any occurs and then our team submit it.

## Dataset Analysis:

We choose the climate data for analyze purpose. We collect the dataset from the kaggle.com to download the dataset. The dataset contains various irrelevant data like noisy using the python libraries like pandas to remove the noisy data and also use the IBM db2 to remove the noisy data using the queries and scripts in SQL section.

rainfall in India 1901-2015.ods - LibreOffice Calc

FileEditViewInsertFormatStylesSheetDataToolsWindowHelp

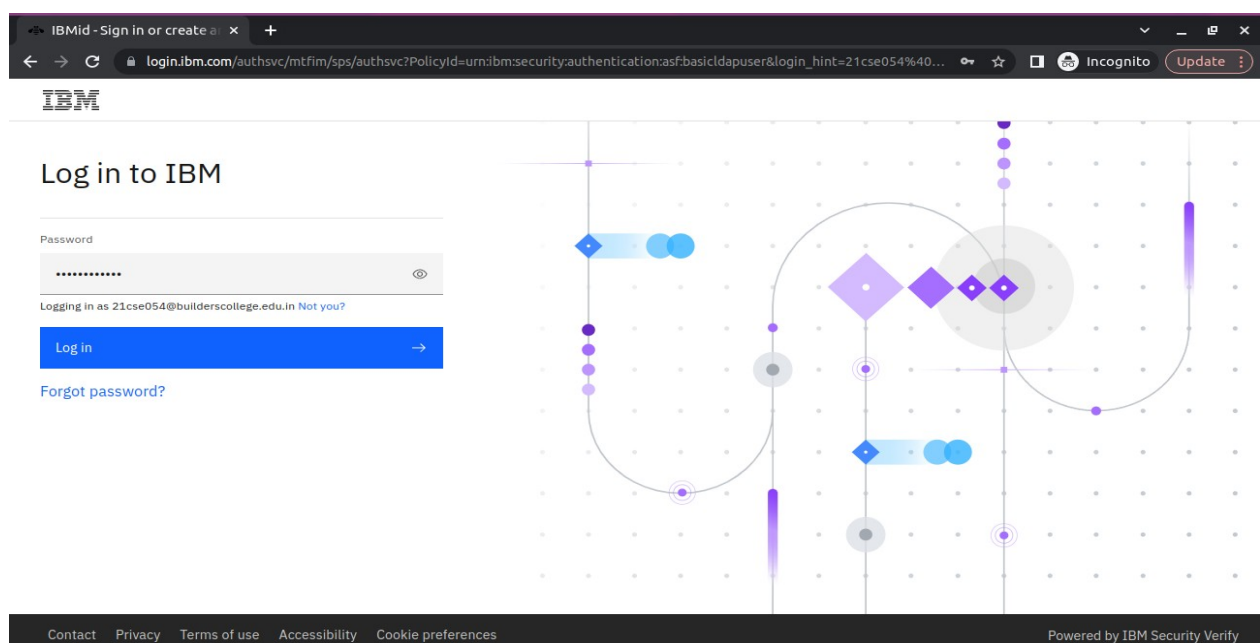
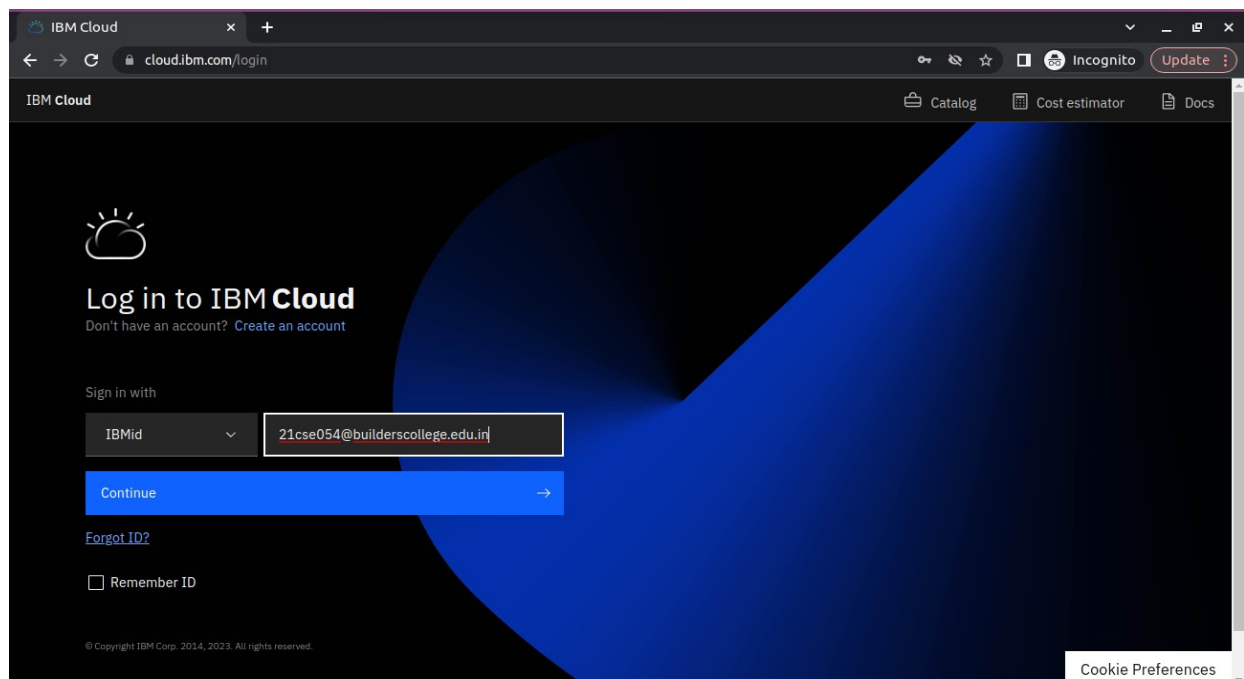
## Setup Database and upload the dataset into db2:

### Steps followed to create IBM Cloud Account.

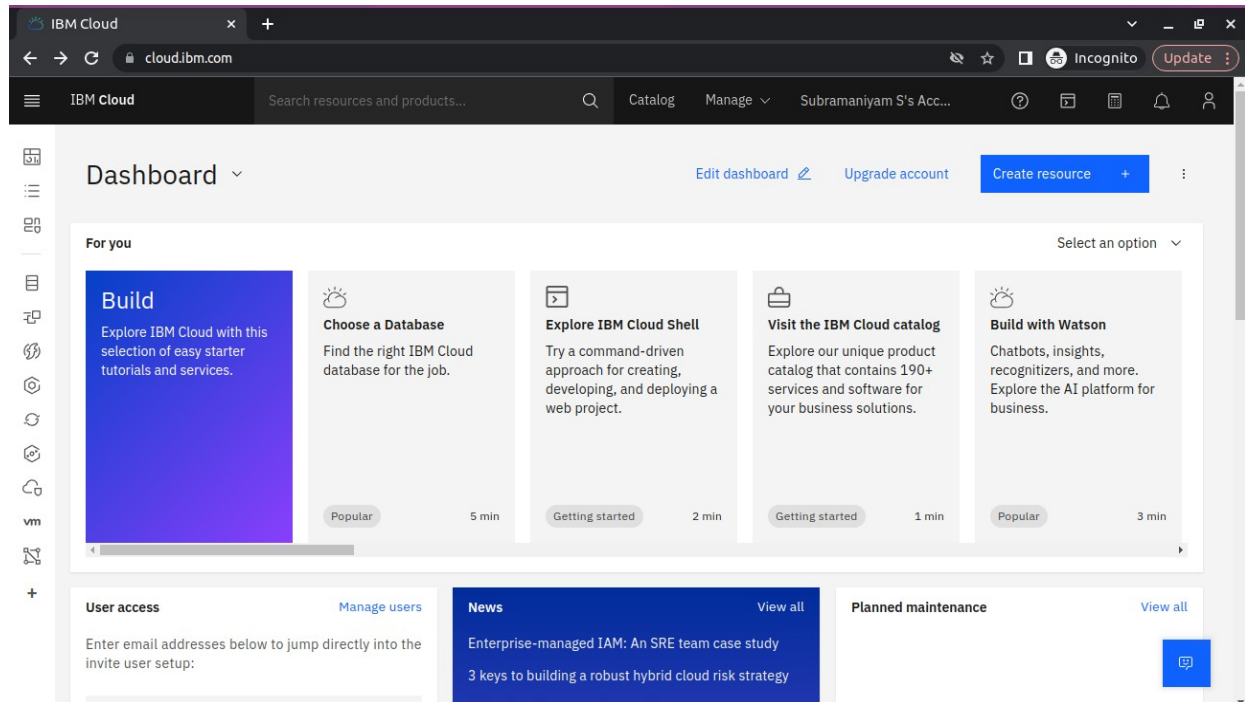
**Step 1** - Go to [myclass.skillup.online](https://myclass.skillup.online) platform and search Introduction to cloud.

**Step 2** - Then Go to Module 1 and click the obtain feature code to obtain the feature code. Then click to activate the account.

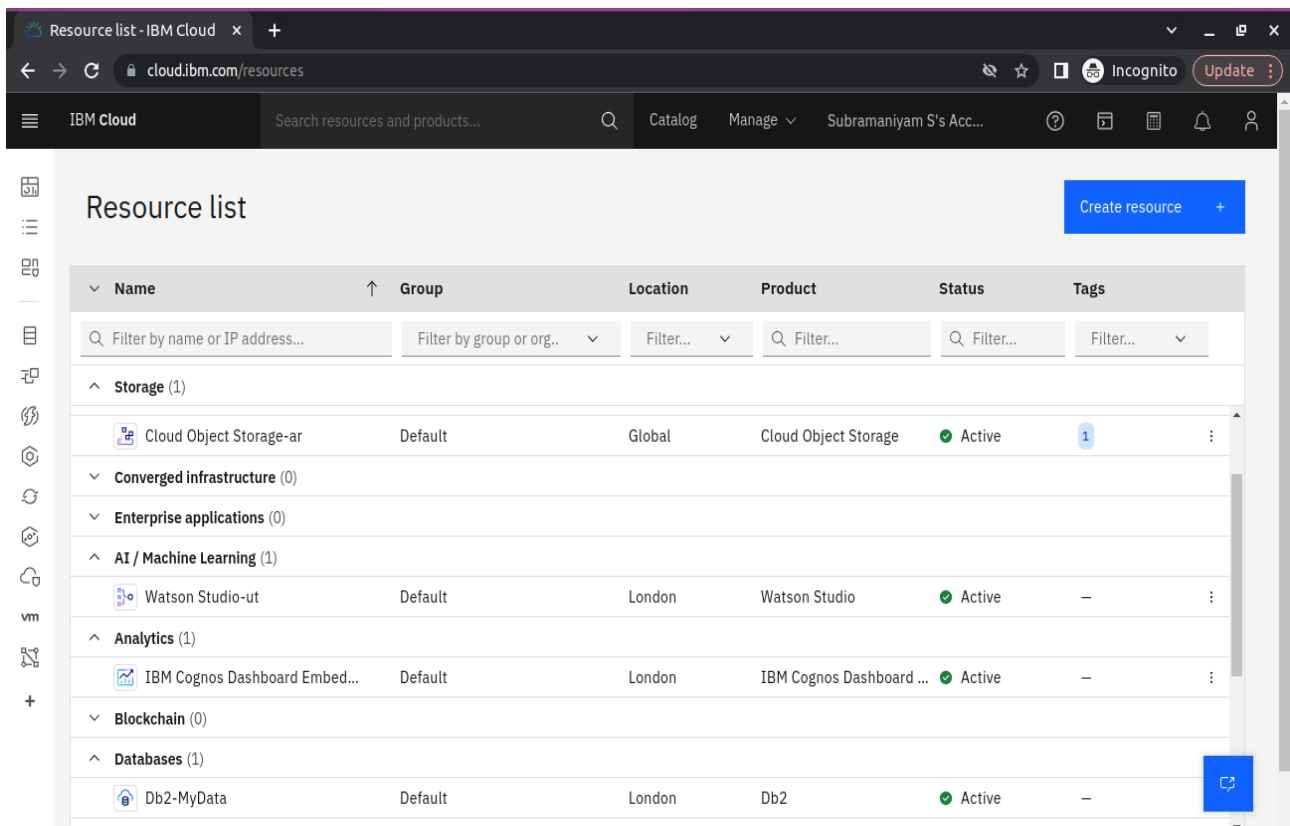
**Step 3** - Then you will be redirected to the [cloud.ibm.com](https://cloud.ibm.com) and then type our registered email id and password.



**Step 4** - Then you will be redirected to the dashboard.

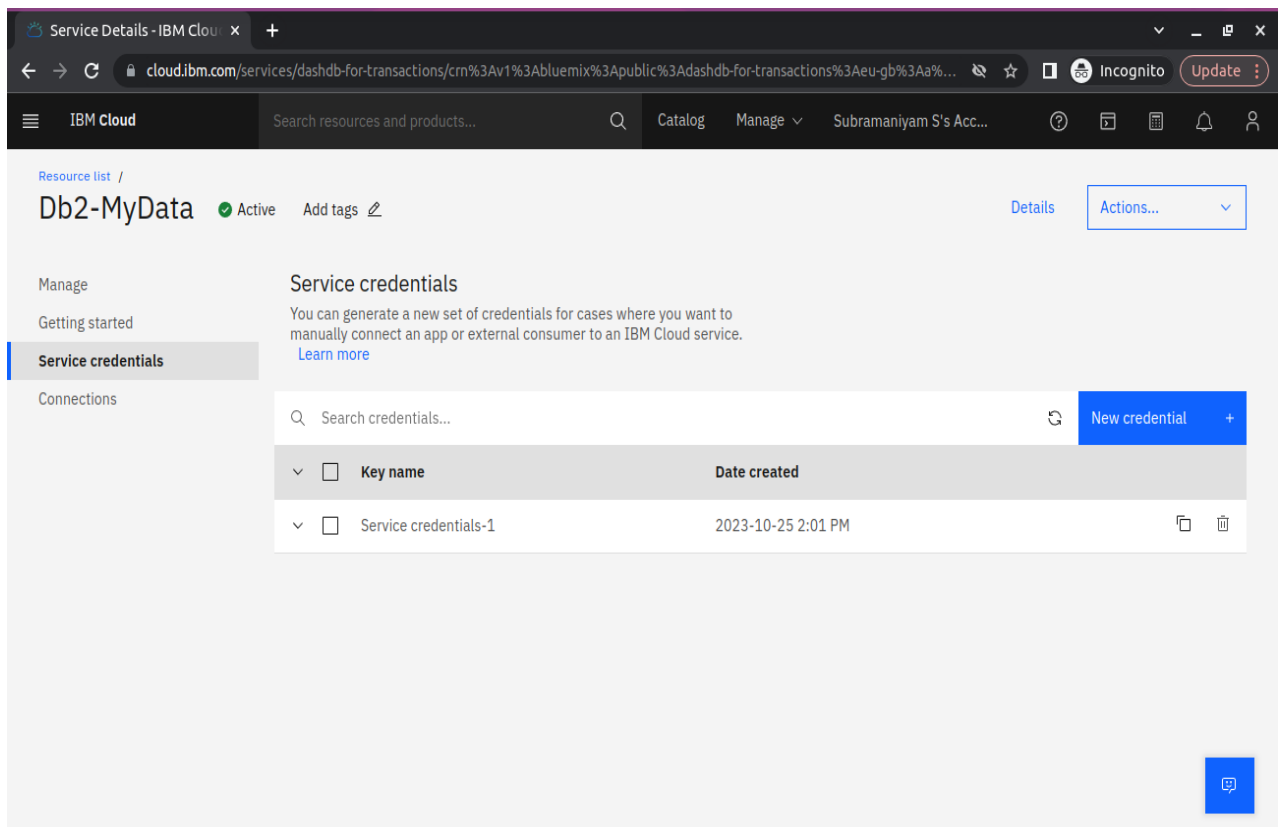
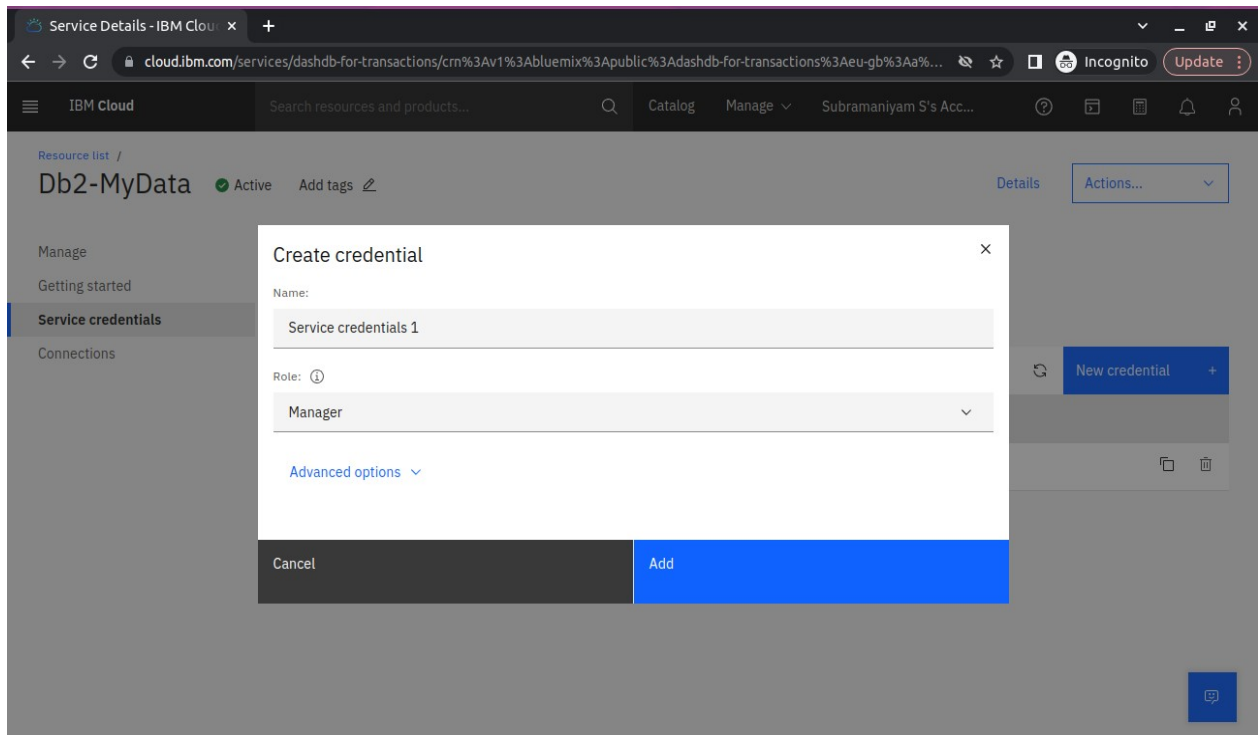


**Step 5** - Then you go to catalog and add db2 to our resources. After successful creation go to resource list and see that.

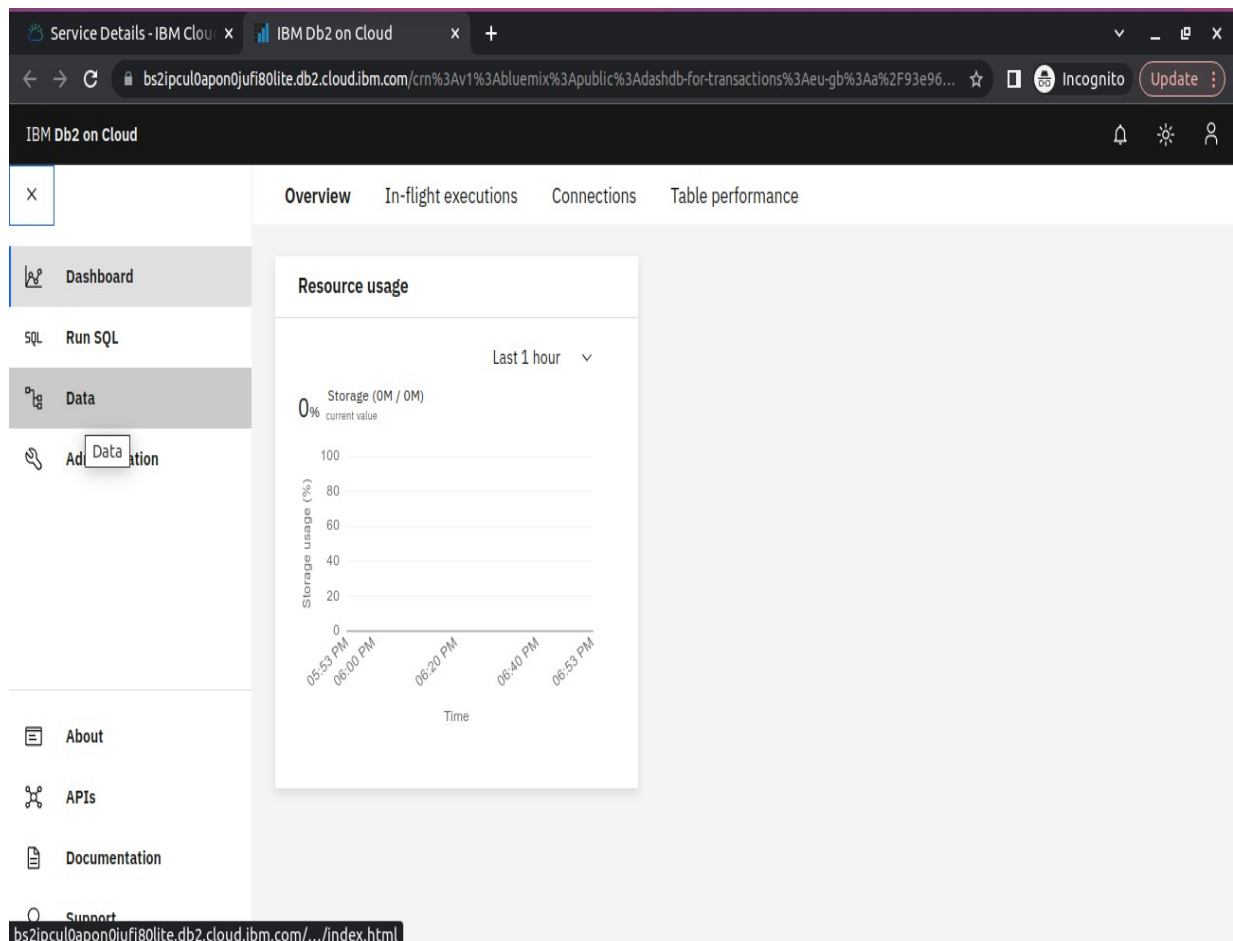
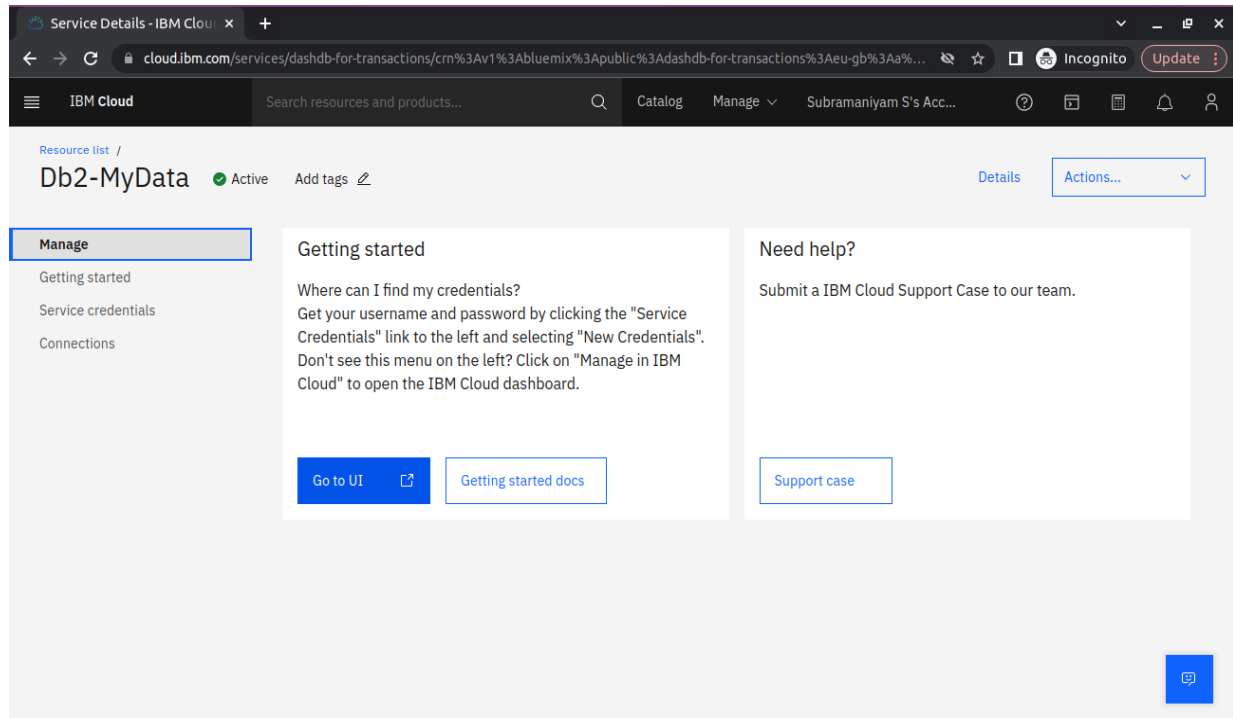


## Data exploration and Analysis:

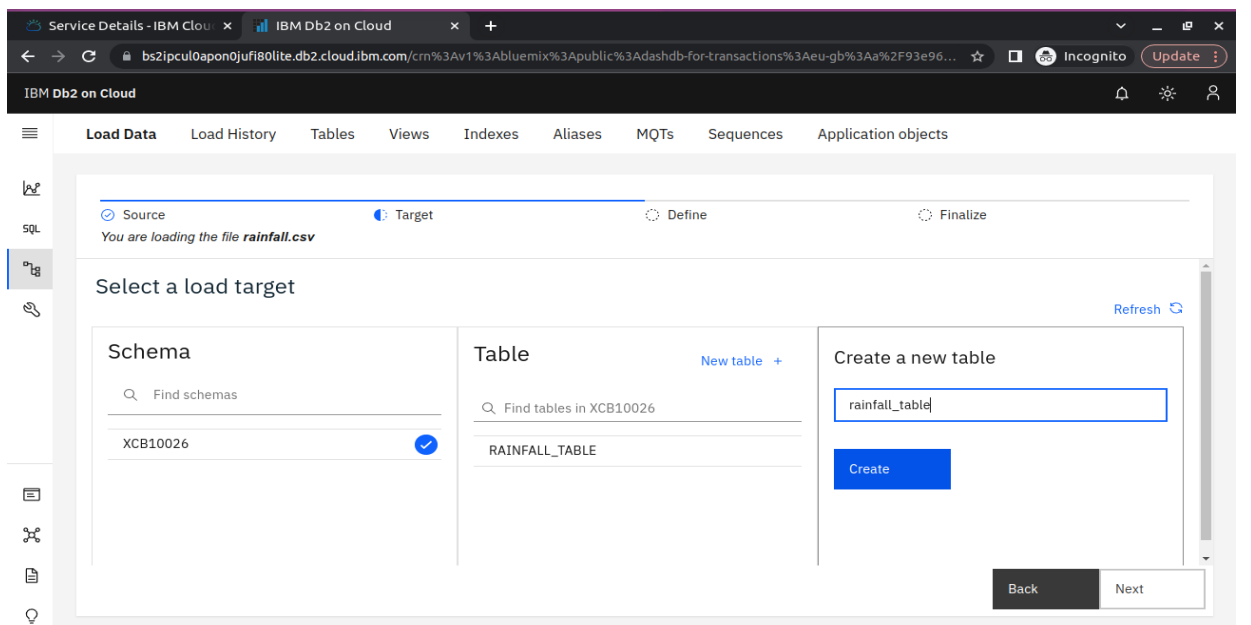
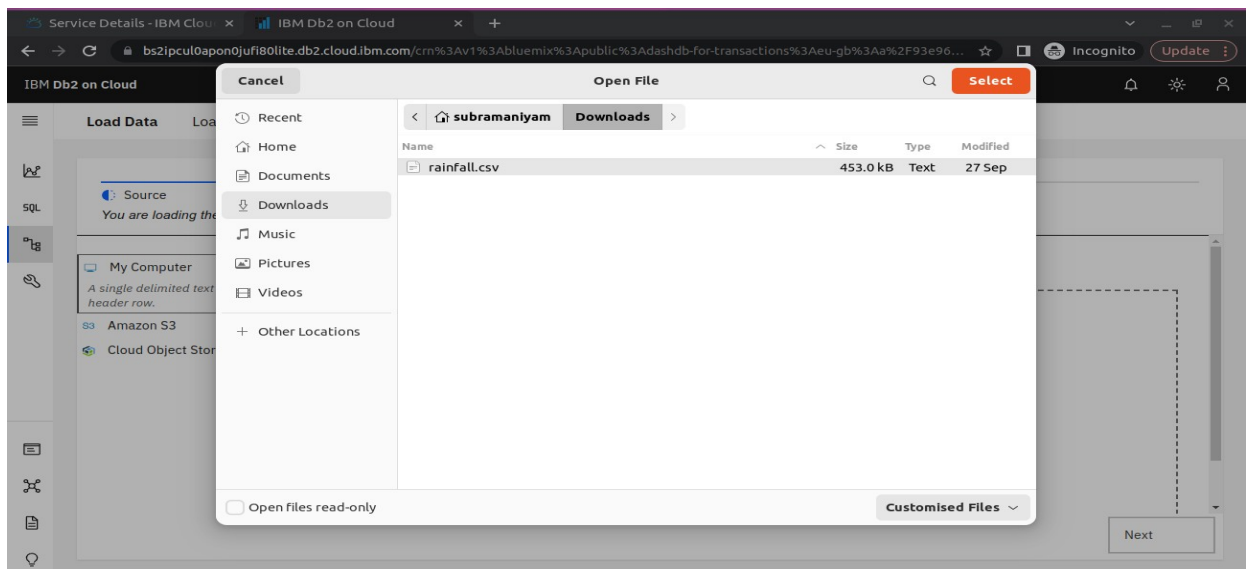
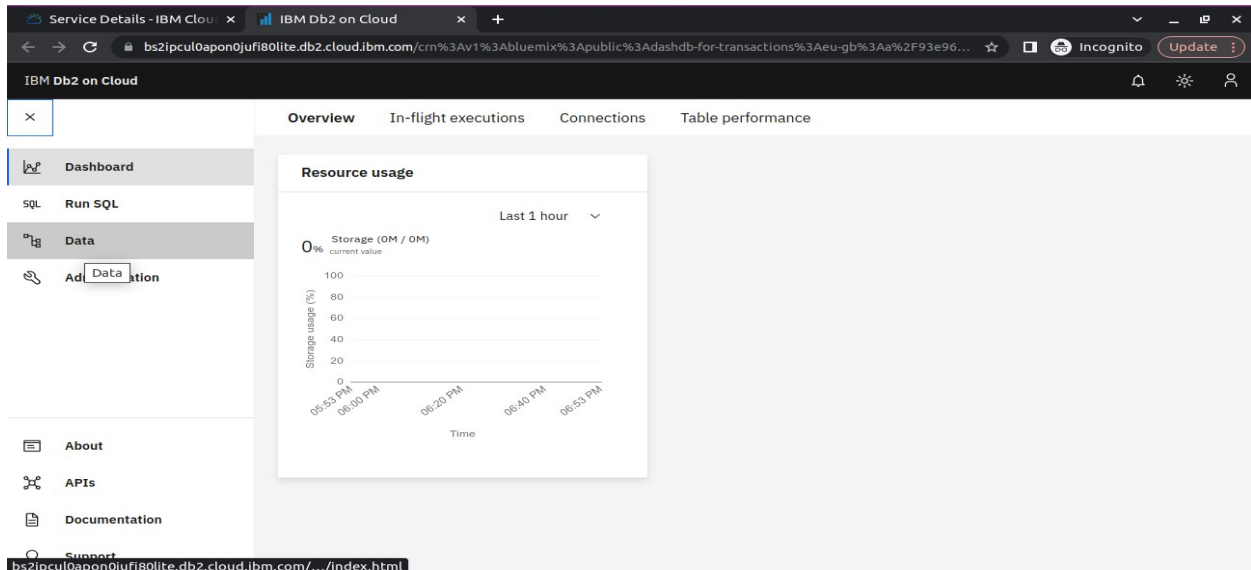
**Step 6** – Then click the database you created in the resource poll and click the service credentials and create new services.

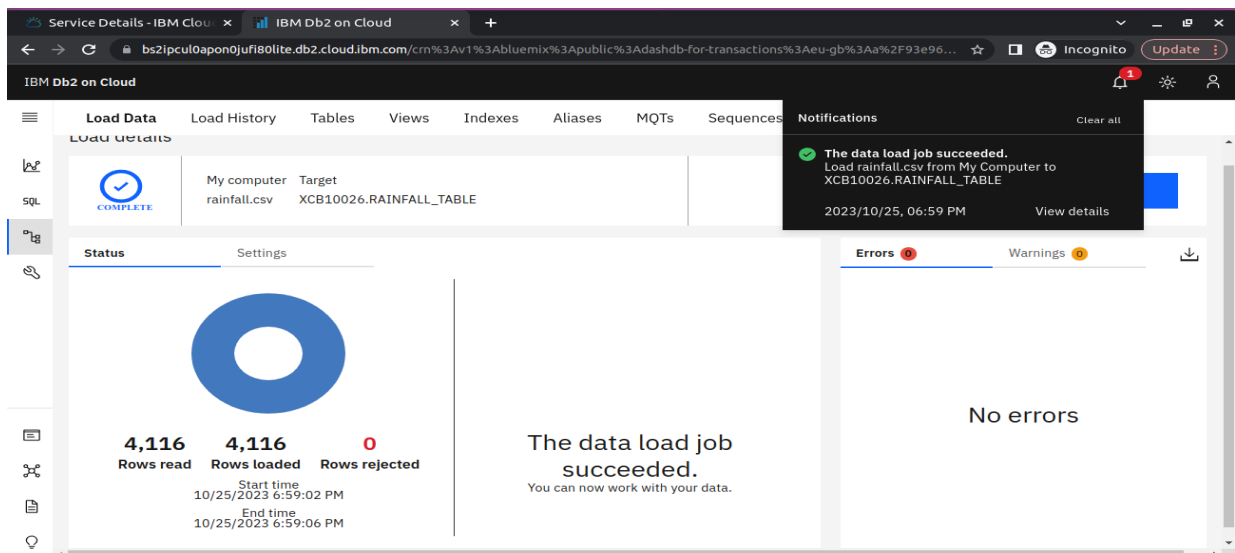


**Step 7** – Then go to manage section and click go to UI.



## Step 8 – Go to data section and upload the dataset that is csv file.





Finally, dataset can be uploaded successfully in db2.

### Step 9- Using SQL queries to analyze the data.

Service Details - IBM Cloud | IBM Db2 on Cloud

bs2ipcul0aponojufi80lite.db2.cloud.ibm.com/crn%3Av1%3Abluemix%3Apublic%3Adashdb-for-transactions%3Aeu-gb%3Aa%2F93e96...

IBM Db2 on Cloud

Data objects | Saved objects

Find objects

XCB10026

Tables | Views | MQTs | Aliases | Nicknames

\*Untitled ...

Syntax assistant

Run selected

1 select subdivision,annual,year from rainfall\_table where(year=2015)

History | Results

Result set 1

Filter table

Total: 72

SUBDIVISION	ANNUAL	YEAR
ANDAMAN & NICOBAR ISLANDS	2904.6	2015
ARUNACHAL PRADESH	2767.5	2015
ASSAM & MEGHALAYA	2470.9	2015
NAGA MANI MIZO TRIPURA	1922.4	2015
SUB HIMALAYAN WEST BENGAL & SIKKIM	2518.6	2015
GANGETIC WEST BENGAL	1530.3	2015

Items per page: 50 | 1-50 of 72 Items

1 | 1 of 2 pages

Again, using the SQL queries to analyze the dataset to find the hidden trends and patterns.

Service Details - IBM Cloud | IBM Db2 on Cloud

bs2ipcul0aponojufi80lite.db2.cloud.ibm.com/crn%3Av1%3Abluemix%3Apublic%3Adashdb-for-transactions%3Aeu-gb%3Aa%2F93e96...

IBM Db2 on Cloud

Data objects | Saved objects

Find objects

XCB10026

Tables | Views | MQTs | Aliases | Nicknames

\*Untitled ...

Syntax assistant

Run selected

1 select subdivision,annual from rainfall\_table where year=2015 order by annual desc LIMIT 5;

History | Results

Result set 1

Filter table

Total: 5

SUBDIVISION	ANNUAL
COASTAL KARNATAKA	3106.0
COASTAL KARNATAKA	3106.0
ANDAMAN & NICOBAR ISLANDS	2904.6
ANDAMAN & NICOBAR ISLANDS	2904.6
ARUNACHAL PRADESH	2767.5

Finally, we analyze the dataset using the IBM db2.

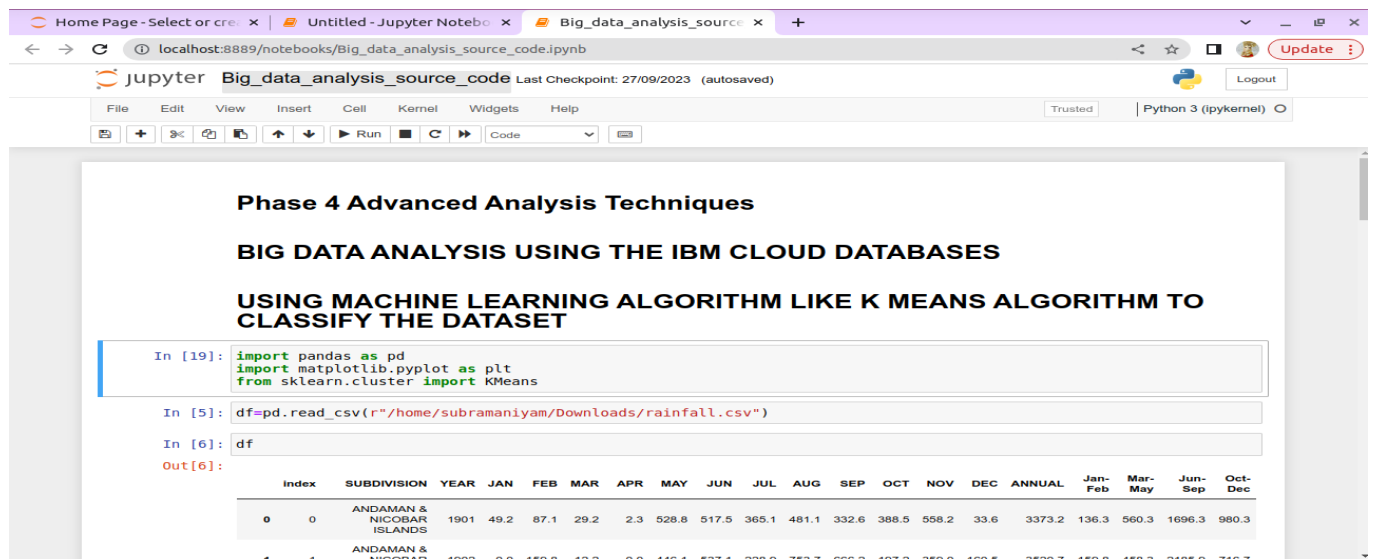


# Advanced Analytics Techniques:

## Follow the below steps for Advanced Analytics Techniques:

**NOTE:** We are going to use the Machine Learning Algorithm like K Means Clustering Algorithm for analysis.

### Step 1 – Import the necessary libraries and the dataset in Jupyter Notebook.



The screenshot shows a Jupyter Notebook interface with the following code and output:

```
In [19]: import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

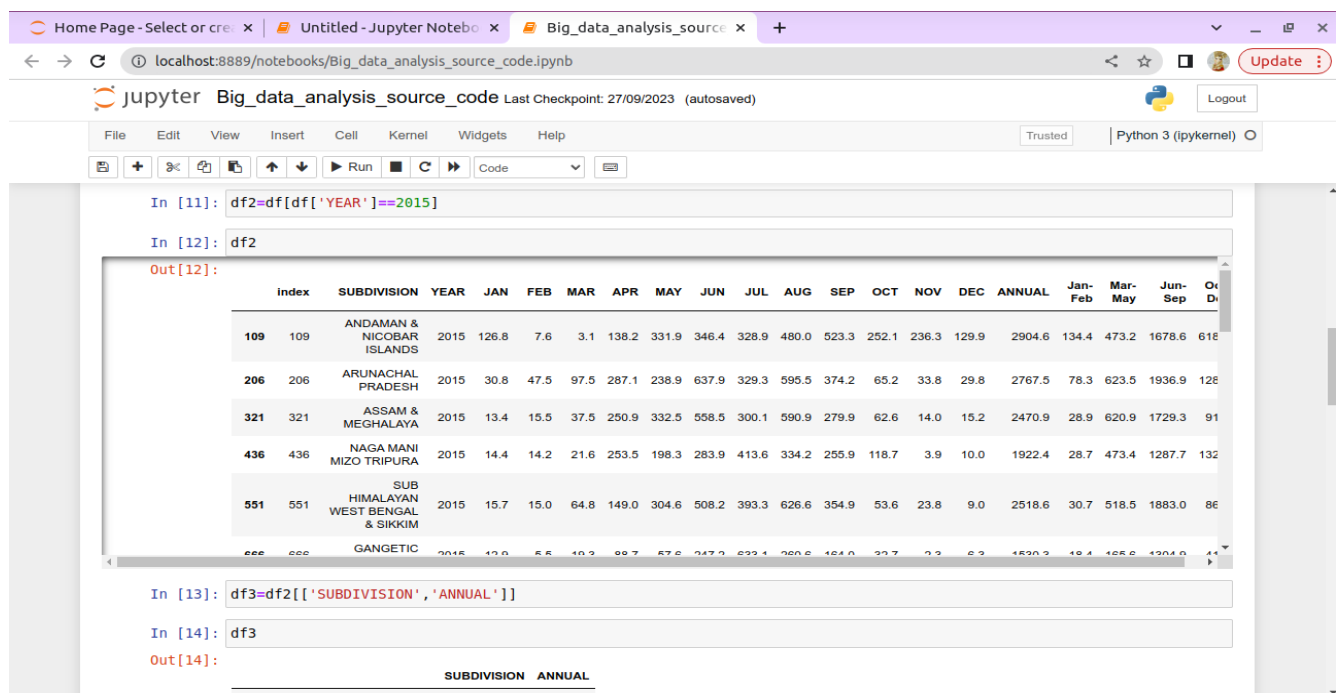
In [5]: df=pd.read_csv(r"/home/subramaniyam/Downloads/rainfall.csv")

In [6]: df
```

Out[6]:

	Index	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec
0	0	ANDAMAN & NICOBAR ISLANDS	1901	49.2	87.1	29.2	2.3	528.8	517.5	365.1	481.1	332.6	388.5	558.2	33.6	3373.2	136.3	560.3	1696.3	980.3
1	1	ANDAMAN & NICOBAR	1902	0.0	159.8	12.2	0.0	446.1	537.1	228.9	753.7	666.2	197.2	359.0	160.5	3520.7	159.8	458.3	2185.9	716.7

### Step 2 – Clean and remove the noisy data in the dataset using python script.



The screenshot shows a Jupyter Notebook interface with the following code and output:

```
In [11]: df2=df[df['YEAR']==2015]

In [12]: df2
```

Out[12]:

	Index	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec
109	109	ANDAMAN & NICOBAR ISLANDS	2015	126.8	7.6	3.1	138.2	331.9	346.4	328.9	480.0	523.3	252.1	236.3	129.9	2904.6	134.4	473.2	1678.6	618.0
206	206	ARUNACHAL PRADESH	2015	30.8	47.5	97.5	287.1	238.9	637.9	329.3	595.5	374.2	65.2	33.8	29.8	2767.5	78.3	623.5	1936.9	126.0
321	321	ASSAM & MEGHALAYA	2015	13.4	15.5	37.5	250.9	332.5	558.5	300.1	590.9	279.9	62.6	14.0	15.2	2470.9	28.9	620.9	1729.3	91.0
436	436	NAGA MANI MIZO TRIPURA	2015	14.4	14.2	21.6	253.5	198.3	283.9	413.6	334.2	255.9	118.7	3.9	10.0	1922.4	28.7	473.4	1287.7	132.0
551	551	SUB HIMALAYAN WEST BENGAL & SIKKIM	2015	15.7	15.0	64.8	149.0	304.6	508.2	393.3	626.6	354.9	53.6	23.8	9.0	2518.6	30.7	518.5	1883.0	86.0
666	666	GANGETIC	2015	11.0	5.5	10.2	88.7	57.6	247.2	632.1	280.6	164.0	32.7	2.3	6.2	1530.2	16.4	165.6	1304.0	11.0

```
In [13]: df3=df2[['SUBDIVISION', 'ANNUAL']]

In [14]: df3
```

Out[14]:

	SUBDIVISION	ANNUAL
109	ANDAMAN & NICOBAR ISLANDS	2904.6

Home Page - Select or create a new notebook | Untitled - Jupyter Notebook | Big\_data\_analysis\_source | +

localhost:8889/notebooks/Big\_data\_analysis\_source\_code.ipynb

Jupyter Big\_data\_analysis\_source\_code Last Checkpoint: 27/09/2023 (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

In [13]: `df3=df2[['SUBDIVISION', 'ANNUAL']]`

In [14]: `df3`

Out[14]:

	SUBDIVISION	ANNUAL
109	ANDAMAN & NICOBAR ISLANDS	2904.6
206	ARUNACHAL PRADESH	2767.5
321	ASSAM & MEGHALAYA	2470.9
436	NAGA MANI MIZO TRIPURA	1922.4
551	SUB HIMALAYAN WEST BENGAL & SIKKIM	2518.6
666	GANGETIC WEST BENGAL	1530.3
781	ORISSA	1210.1
896	JHARKHAND	1081.8
1011	BIHAR	872.7
1126	EAST UTTAR PRADESH	603.3
1241	WEST UTTAR PRADESH	582.7
1356	UTTARAKHAND	1247.6
1471	HARYANA DELHI & CHANDIGARH	435.3
1586	PUNJAB	510.8
1701	HIMACHAL PRADESH	1210.5
1816	JAMMU & KASHMIR	1572.8
1931	WEST RAJASTHAN	458.4

### Step 3 – Store the necessary values into the empty array for plotting purpose.

Home Page - Select or create a new notebook | Untitled - Jupyter Notebook | Big\_data\_analysis\_source | +

localhost:8889/notebooks/Big\_data\_analysis\_source\_code.ipynb

Jupyter Big\_data\_analysis\_source\_code Last Checkpoint: 27/09/2023 (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

In [61]:

```
df4=df3['SUBDIVISION']
df5=df3['ANNUAL']
states=[]
annual_rainfall_values=[]
for i in range(0,len(df4),1):
    states.append(i)
    annual_rainfall_values.append(df5.iloc[i])
print(states)
print(annual_rainfall_values)
```

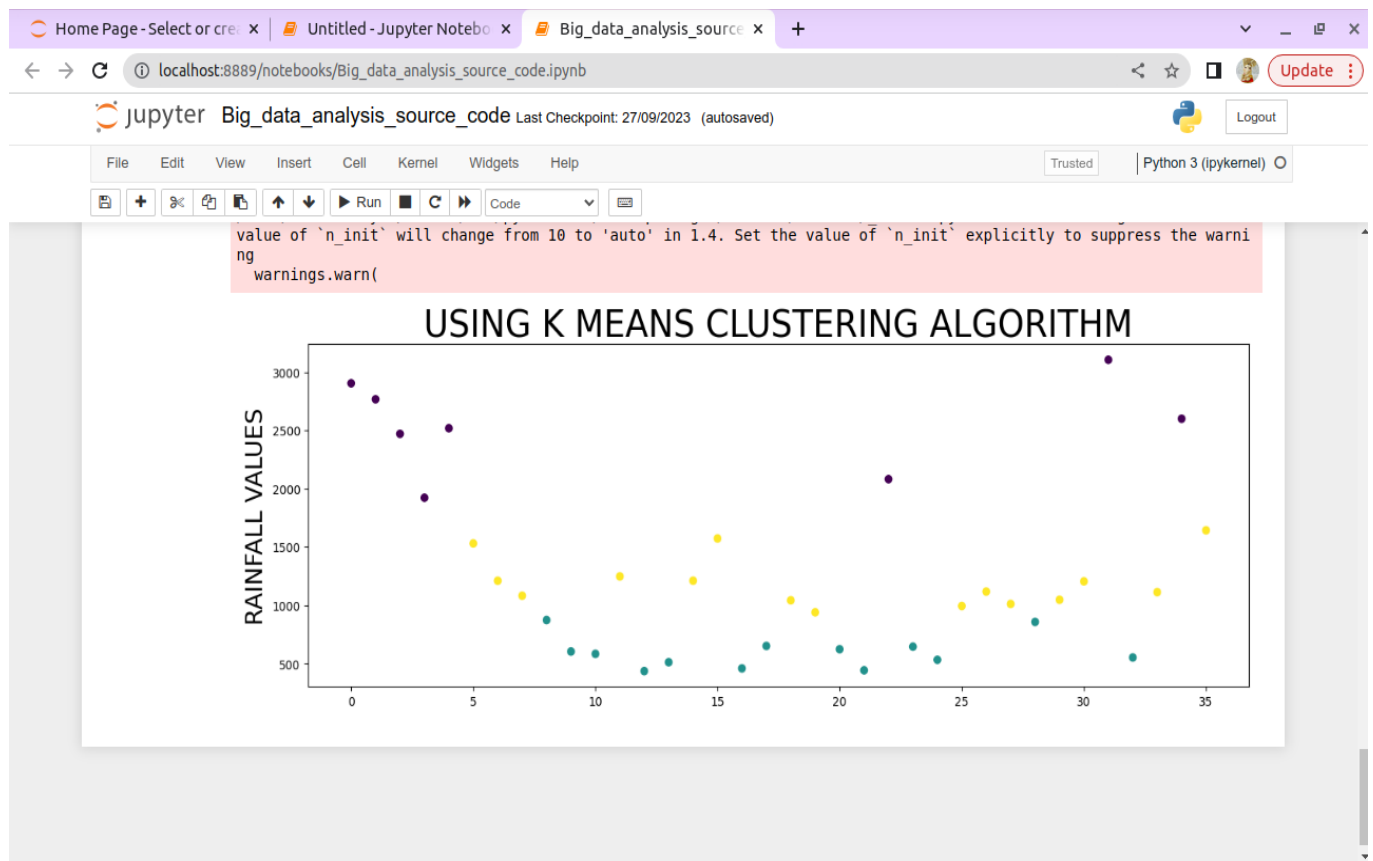
```
[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35]
[2904.6, 2767.5, 2470.9, 1922.4, 2518.6, 1530.3, 1210.1, 1081.8, 872.7, 603.3, 582.7, 1247.6, 435.3, 510.8, 1210.5, 1572.8, 458.4, 650.7, 1042.3, 939.2, 622.9, 441.7, 2082.0, 644.5, 532.2, 993.8, 1117.6, 1010.9, 857.3, 1047.1, 1204.6, 3106.0, 551.9, 1112.5, 2600.6, 1642.9]
```

In [67]:

```
whole_data=list(zip(states,annual_rainfall_values))
kmeans = KMeans(n_clusters=3)
kmeans.fit(whole_data)
width1 = 15.
height1 = 5.
width,height1 = (width1, height1)
plt.figure(figsize=(width,height1))
plt.title("USING K MEANS CLUSTERING ALGORITHM",fontsize=30)
plt.xlabel("RAINFALL VALUES",fontsize=20)
plt.scatter(states,annual_rainfall_values , c=kmeans.labels_)
plt.show()
```

/home/subramaniyam/.local/lib/python3.10/site-packages/sklearn/cluster/\_kmeans.py:870: FutureWarning: The default value of 'n\_init' will change from 10 to 'auto' in 1.4. Set the value of 'n\_init' explicitly to suppress the warning

**Step 4** – After storing the values in array using K Means Clustering Algorithm to plot the graph and analyze the results.



**Visualization:** Create visualizations to showcase the analysis results. Use tools like Matplotlib, Plotly, or IBM Watson Studio for creating graphs and charts.

**Follow the below steps for Advanced Analytics Techniques:**

**NOTE:** We are going to use the IBM Watson Studio for creating graphs and charts.

**Step 1** – Open our cloud account and create the IBM Watson Studio then go to the Resource Pool and click Artificial Intelligence and Machine Learning and choose IBM Watson Studio.

Activities Google Chrome Oct 25 15:08

Resource x Service D x Data Refi x IBM Db2 x SQL SELE x Upload ar x Select Cas x IBM WAT x +

eu-gb.dataplatform.cloud.ibm.com/shaper?project\_id=64804e62-d699-401f-bf9e-24b7e951c21b&dataset\_id=a58c1ec8-e... Incognito Update

IBM Cloud Pak for Data Search in your workspaces Upgrade Subramaniyam S's Account London SS

Projects / rainfall\_visualization / rainfall-1.csv / Data Refinery

Steps (3)

Data source: rainfall-1.csv

1. Convert column type: Automatically converted one or more columns to inferred data types. Strings that are converted to decimal use a dot (.) for the decimal symbol. Auto-generated

2. Custom code: select('SUBDIVISION','YEAR','ANNUAL')

3. Custom code: filter('YEAR' == 2015) Just added

New step +

Use a code template to add a step

Data Profile Visualizations

	SUBDIVISI...	YEAR	ANNUAL
	String	Integer	Decimal
1	ANDAMAN & NICO...	2015	2904.6
2	ARUNACHAL PRAD...	2015	2767.5
3	ASSAM & MEGHALA...	2015	2470.9
4	NAGA MANI MIZO T...	2015	1922.4
5	SUB HIMALAYAN W...	2015	2518.6
6	GANGETIC WEST B...	2015	1530.3
7	ORISSA	2015	1210.1
8	JHARKHAND	2015	1081.8
9	BIHAR	2015	872.7
10	EAST UTTAR PRAD...	2015	603.3
11	WEST UTTAR PRAD...	2015	582.7
	UTTARAKHAND	2015	1247.6

Configure Viewing: 36 rows, 3 columns Full data set: 4116 rows, 20 columns

**Step 2 – Load the Dataset and put some queries to refine the data for our visualization.**

Screenshots - Google Driv x Resource list - IBM Cloud x +

cloud.ibm.com/resources

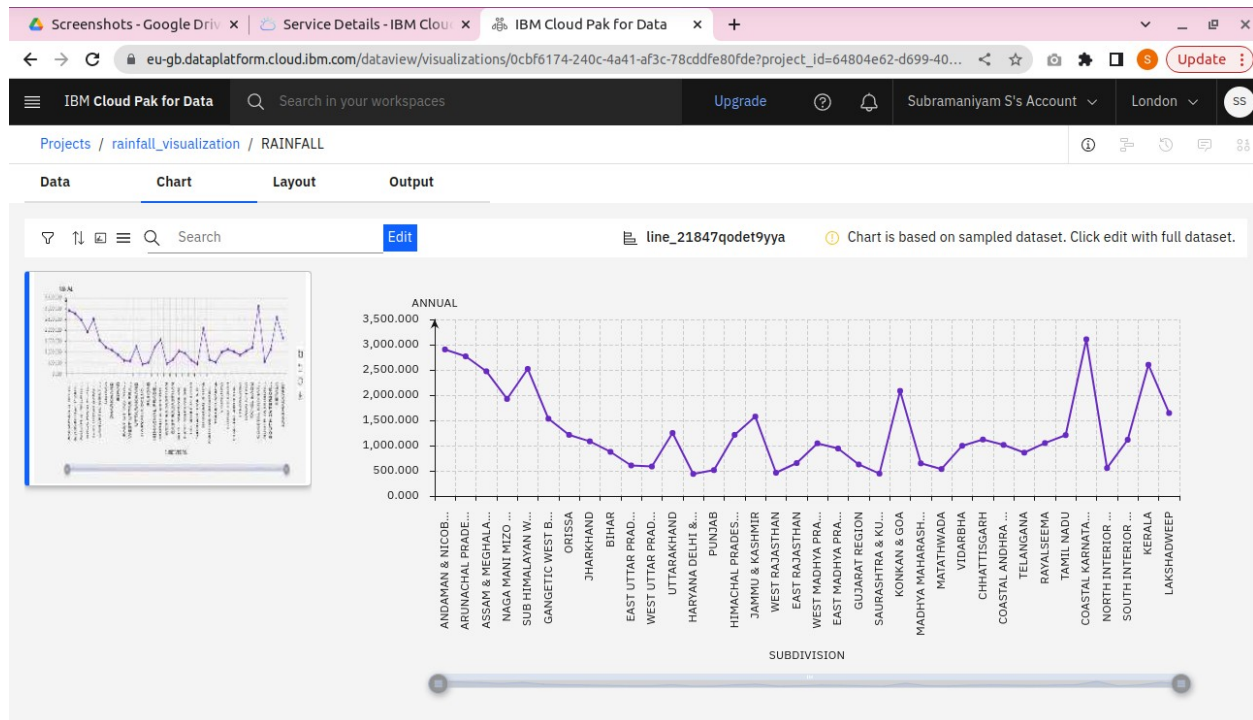
IBM Cloud Search resources and products... Catalog Manage Subramaniyam S's Acc...

## Resource list

Create resource +

Name	Group	Location	Product	Status	Tags
Filter by name or IP address... Filter by group or org... Filter... Filter... Filter... Filter...					
Networking (0)					
Storage (1)					
Cloud Object Storage-ar	Default	Global	Cloud Object Storage	Active	1
Converged infrastructure (0)					
Enterprise applications (0)					
AI / Machine Learning (1)					
Watson Studio-ut	Default	London	Watson Studio	Active	—
Analytics (1+)					
Blockchain (0)					
Databases (1+)					
Developer tools (0)					

### Step 3 – Finally, using the refine script to perform the visualization.



### Conclusion:

In this project, we clearly understand the big data analytics topics and how to perform analytics work using IBM db2 and visualization works using IBM Watson Studio and Python.