SYMBIOSIS INSTITUTE OF TECHNOLOGY, NAGPUR

Water Reservoir EDA

Data Science Report

Submitted To:

Dr. Piyush Chauhan

Associate Professor

Submitted By:
Gautam Sukhani, 22070521059

Sem: 6, Section: A

August 2025

Index Table-

1. Introduction

This report presents a comprehensive analysis of the provided reservoir water data, which spans from 2011 to 2021. The primary objective is to transform raw data into valuable insights regarding reservoir usage, water storage trends, and regional performance. By following a structured approach from cleaning and preparing the data to conducting detailed analyses we aim to uncover key patterns and provide a clear, easy-to-understand narrative of the findings. This document will serve as a foundational resource for understanding the current state of water storage across the analyzed regions.

2. Dataset Overview

	A	B	C	D	E	F	G	H	I	J
1	Country	State	District	Year	Month	Reservoir Basin Name	Reservoir Name	Full Reservoir Lev	Reservoir W	Reservoir
2	India	Andhra Pradesh	Anantapur	Calendar Year (Jan - Dec), 2021	May, 2021	PENNAR	GODDAMVARIPALLI SR3	0.01		
3	India	Andhra Pradesh	Anantapur	Calendar Year (Jan - Dec), 2021	May, 2021	PENNAR	CHITRAVATI BALANCING RESERVOIR	0.28	297.92	0.01
4	India	Andhra Pradesh	Anantapur	Calendar Year (Jan - Dec), 2021	May, 2021	PENNAR	MID PENNAR RESERVOIR	0.15	356.94	0
5	India	Andhra Pradesh	Anantapur	Calendar Year (Jan - Dec), 2021	May, 2021	PENNAR	PENNA AHOBILAM BALANCING RESERVOIR	0.31	431.78	0.01
6	India	Andhra Pradesh	Anantapur	Calendar Year (Jan - Dec), 2021	May, 2021	PENNAR	PENNAR KUMUDWATHI PROJECT (ANICUT)			
7	India	Andhra Pradesh	Anantapur	Calendar Year (Jan - Dec), 2021	May, 2021	PENNAR	UPPER PENNAR	0.05	0	0
8	India	Andhra Pradesh	Anantapur	Calendar Year (Jan - Dec), 2021	May, 2021	PENNAR	CHAGALLU BALANCING RESERVOIR	0.05	264.08	0
9	India	Andhra Pradesh	Anantapur	Calendar Year (Jan - Dec), 2021	May, 2021	KRISHNA	BHAIRAVANITHIPPA PROJECT	0.06	0	0
10	India	Andhra Pradesh	Anantapur	Calendar Year (Jan - Dec), 2021	May, 2021	PENNAR	Cherlopalli Reservoir	0.05		
11	India	Andhra Pradesh	Anantapur	Calendar Year (Jan - Dec), 2021	May, 2021	PENNAR	GODDUMARRI ANICUT	0.01		
12	India	Andhra Pradesh	Anantapur	Calendar Year (Jan - Dec), 2021	May, 2021	PENNAR	CHENNARAYA SWAMI GUDI PROJECT	0	580.95	0
13	India	Andhra Pradesh	Anantapur	Calendar Year (Jan - Dec), 2021	May, 2021	PENNAR	YELLANUR SR2	0.01		
14	India	Andhra Pradesh	Anantapur	Calendar Year (Jan - Dec), 2021	May, 2021	PENNAR	YOGI VEMANA (MADDILERU) RESERVOIR	0.03	379.04	0
15	India	Andhra Pradesh	Anantapur	Calendar Year (Jan - Dec), 2021	May, 2021	PENNAR	GOLLAPALLI RESERVOIR	0.05		

- Entries: 53,625
- Columns: 10
- Data Coverage: Reservoir data across multiple districts and states in India.

Columns Overview

1. Country: All entries are from India.
2. State: Includes 20 states.
3. District: 117 districts are represented.
4. Year: Contains yearly data (e.g. Calendar Year 2020, 2021, etc. from 2011 to 2021 [10 years]).
5. Month: Monthly data points (125 unique values).
6. Reservoir Basin Name: 19 unique river basins (e.g., PENNAR, GANGA).
7. Reservoir Name: 429 unique reservoirs.
8. Full Reservoir Capacity (BCM): Storage capacity in billion cubic meters
9. Reservoir Water Level (Meters): Actual water level
10. Reservoir Water Storage (BCM): Actual water storage

Value of the Dataset

- Water Resource Management: Provides historical and regional reservoir data crucial for water resource planning, drought management, and flood control.
- Policy Making: Enables policymakers to assess state-wise reservoir performance and plan interventions.
- Environmental Studies: Helps analyze climate change effects on water levels and storage.
- Data Science Use: Suitable for predictive modeling (e.g., forecasting reservoir levels), clustering, and trend analysis.

This dataset is as it:

- Covers a large temporal (year, month) and spatial (state, district, reservoir) range.
- Contains key hydrological indicators for understanding water availability.
- Supports decision-making in agriculture, hydropower, and urban water supply sectors.

3. Data Cleaning and Preparation

The initial dataset, consisting of 53,625 records and ten columns, required significant preparation before any meaningful analysis could be conducted. The first step was to streamline the dataset by renaming the columns to more concise and understandable names. This made the subsequent code easier to read and work with.

```
print("Duplicate rows:", df.duplicated().sum())
```

```
df.rename(columns={
    'Full Reservoir Level (Frl) Capacity (UOM:BCM(BillionCubicMeter)), Scaling Factor:1': 'Full_Reservoir_Capacity_BCM',
    'Reservoir Water Level (UOM:M(Meter)), Scaling Factor:1': 'Reservoir_Water_Level_M',
    'Reservoir Water Storage (UOM:BCM(BillionCubicMeter)), Scaling Factor:1': 'Reservoir_Water_Storage_BCM'
}, inplace=True)


df.columns
```

A crucial part of the process was addressing missing values. Several columns, including Reservoir Water Level and Reservoir Water Storage, had a substantial number of missing entries, nearly 50% of their total values. To handle this, the missing values in these numerical columns were imputed using the median of their respective columns. This method was also applied to the Full Reservoir Level Capacity column, which had fewer missing values.

```
df['Reservoir_Water_Level_M'] = df['Reservoir_Water_Level_M'].fillna(df['Reservoir_Water_Level_M'].median())
df['Reservoir_Water_Storage_BCM'] = df['Reservoir_Water_Storage_BCM'].fillna(df['Reservoir_Water_Storage_BCM'].median())
df['Full_Reservoir_Capacity_BCM'] = df['Full_Reservoir_Capacity_BCM'].fillna(df['Full_Reservoir_Capacity_BCM'].median())
```

4. Preprocessing / Feature Engineering

Following the handling of missing data, the dataset's data types were reviewed and adjusted. The `Year` and `Month` columns were initially stored as objects, which is not ideal for numerical or time-based analysis. The `Year` was extracted as an integer, and the `Month` was converted to a proper datetime format. This conversion was essential for any future time-series analysis.

```
df['Year'] = df['Year'].str.extract(r'(\d{4})').astype(int)

df['Month'] = pd.to_datetime(df['Month'], format='%b, %Y',
errors='coerce')

df[['Year', 'Month']].head()
```

After the cleaning phase was complete, a new feature called `Storage_Efficiency_%` was engineered. This metric was calculated by dividing the current water storage by the full reservoir capacity and multiplying the result by 100. This new column provides a clear measure of how full each reservoir is, which is a valuable addition for further analysis.

$$\text{Storage Efficiency}(\%) = \left(\frac{\text{Reservoir Water Storage(BCM)}}{\text{Full Reservoir Capacity(BCM)}}\right) \times 100$$

```
df['Storage_Efficiency_%'] = (df['Reservoir_Water_Storage_BCM'] /
df['Full_Reservoir_Capacity_BCM']) * 100
```

5. Descriptive Statistics

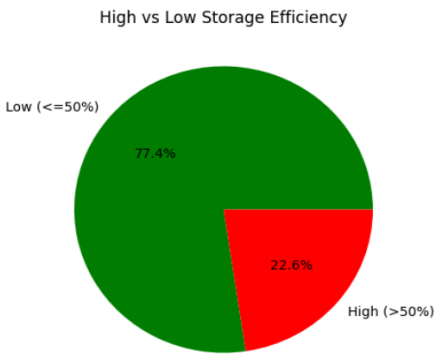
Following the data cleaning and preparation, a descriptive summary was generated to understand the central tendencies, dispersion, and shape of the numerical columns in the dataset. This provided a foundational overview of the data before more in-depth analysis.

	Year	Month	Full_Reservoir_Capacity_BCM	Reservoir_Water_Level_M	Reservoir_Water_Storage_BCM
count	53625.000000	4719	53625.000000	53625.000000	53625.000000
mean	2015.720000	2016-04-30 15:16:21.818181632	36.026746	196.888303	0.195662
min	2011.000000	2011-05-01 00:00:00	0.000000	0.000000	0.000000
25%	2013.000000	2013-05-01 00:00:00	0.100000	139.190000	0.020000
50%	2016.000000	2016-05-01 00:00:00	1.785000	143.750000	0.020000
75%	2018.000000	2019-05-01 00:00:00	8.840000	151.180000	0.020000
max	2021.000000	2021-05-01 00:00:00	7414.290000	1002.790000	9.720000
std	3.013597	NaN	367.648471	172.001115	0.649604

- Average Water Storage: The mean value for Reservoir Water Storage was approximately 0.196 BCM (Billion Cubic Meters). This indicates the typical amount of water held in the reservoirs across the dataset.
- Full Capacity: The average Full Reservoir Capacity was much higher, at around 36.03 BCM. The significant difference between the average storage and the average capacity suggests that, on average, the reservoirs are not filled to their maximum potential.
- Correlation Analysis: A correlation analysis was performed to understand the relationship between a reservoir's full capacity and its actual water storage. The analysis revealed a low correlation coefficient of 0.107 them. This weak positive relationship suggests that a reservoir's maximum capacity is not a strong predictor of its current water storage level. Other factors, such as seasonal variations, local rainfall, or consumption patterns, likely play a more significant role.

State	Reservoir_Water_Storage_BCM	Full_Reservoir_Capacity_BCM	Storage_Efficiency_%
Himachal Pradesh	2.134027	4.156667	45.758237
Madhya Pradesh	1.750222	3.155556	46.676749
Punjab	1.221920	2.340000	52.218803
Odisha	1.066460	1.918750	61.209765
Uttarakhand	0.863733	1.693333	42.573471

```
Top 5 States by Average Storage:
 State
Himachal Pradesh    2.134027
Madhya Pradesh      1.750222
Punjab              1.221920
Odisha              1.066460
Uttarakhand         0.863733
Name: Reservoir_Water_Storage_BCM, dtype: float64
Correlation between Capacity & Storage: 0.10680816607770267
```



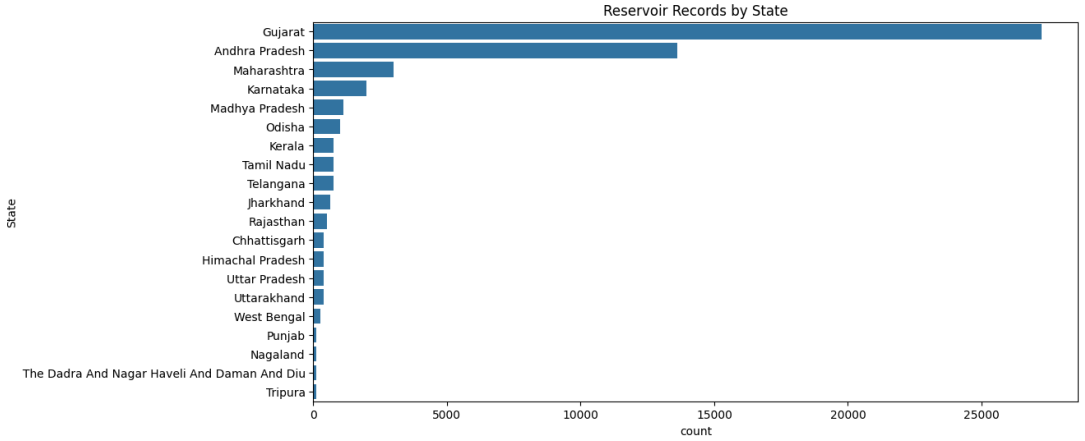
Further analysis was conducted by grouping the data by state to identify regional differences in water storage. The results highlighted the top five states with the highest average reservoir water storage:

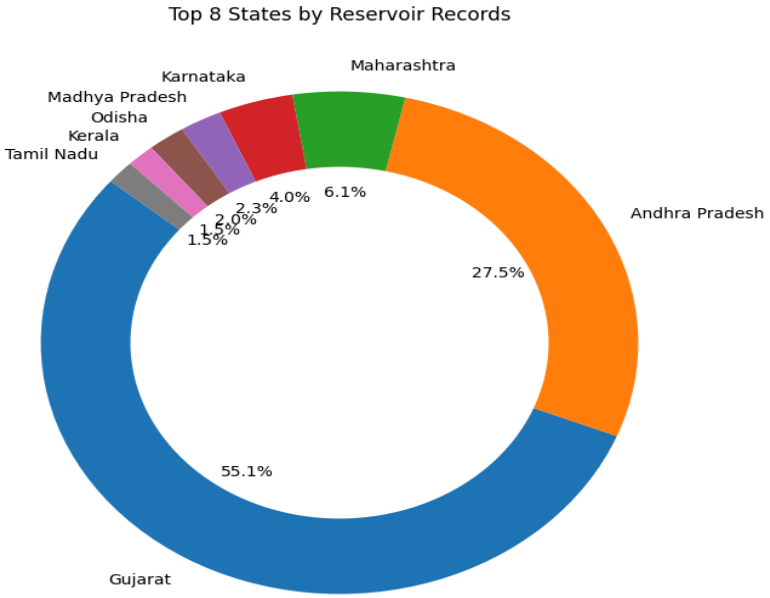
Himachal Pradesh, Madhya Pradesh, Punjab, Odisha, Uttarakhand

This state-level breakdown provides valuable insight into which regions have the largest water reserves on average and their overall efficiency.

6. Univariate Analysis

1. Reservoir Records by State





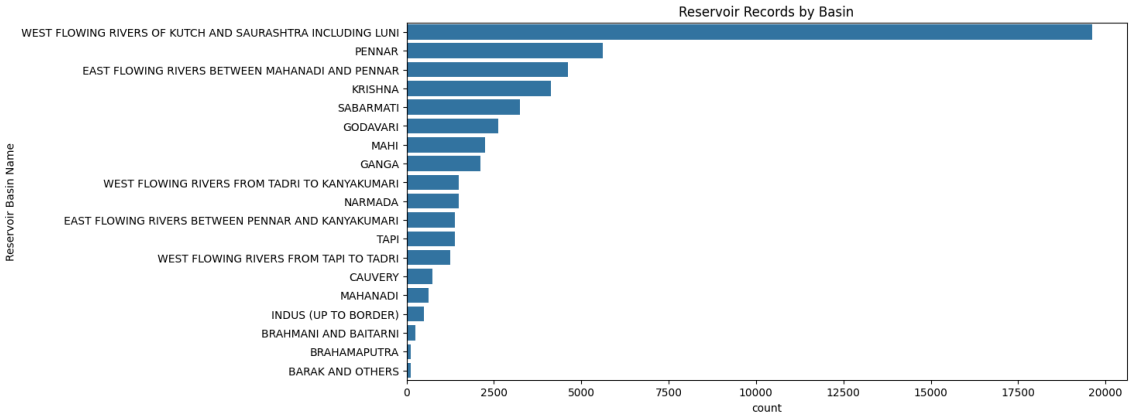
Key Observations:

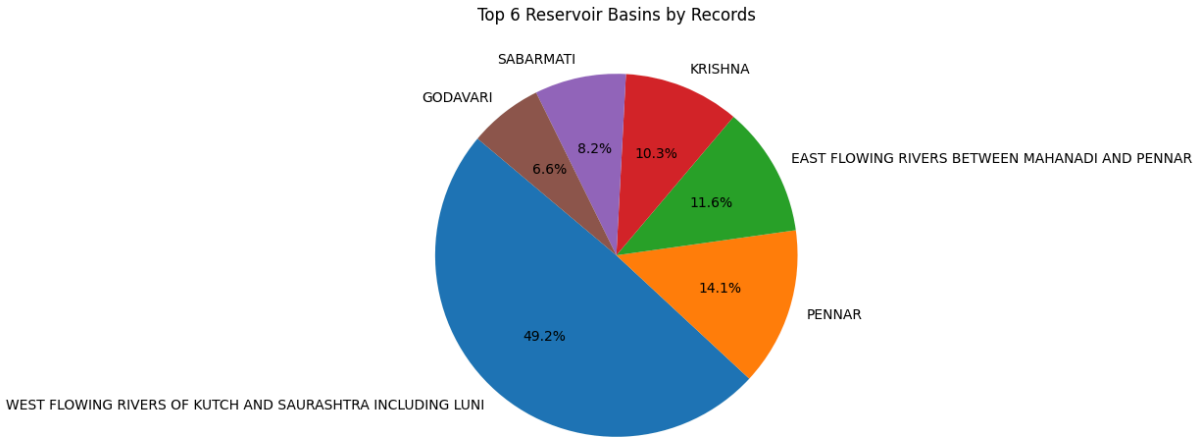
- Gujarat has the highest number of records by a significant margin, with over 25,000 entries and 55%.
- Andhra Pradesh follows with the second-highest count, having more than 10,000 records and over 25%.
- The vast majority of other states have a relatively low number of records, all under 5,000. Many states, such as Punjab, Nagaland, and Tripura, have very few records.

Conclusion:

The primary conclusion from this graph is that the dataset is highly imbalanced with respect to the states. Gujarat and Andhra Pradesh are heavily overrepresented, while most other states are underrepresented. This imbalance is important to note, as it could potentially skew any analysis that does not account for the unequal distribution of data.

2. <u>Reservoir Records by Basin</u>





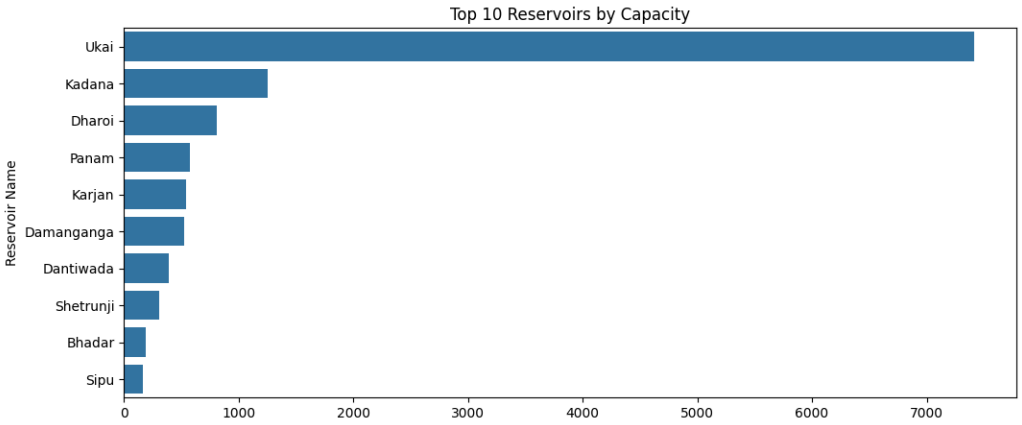
Key Observations:

- Krishna Basin has the highest reservoir count dominating the dataset.
- Godavari and Cauvery basins follow as distant second and third.
- Most basins (e.g., Mahi, Tapi, Brahmani & Baitarni, Barak) have very low representation
- West-flowing rivers (e.g., Tadri to Kanyakumari, Tapi to Tadri) show moderate counts while east-flowing rivers (e.g., Mahanadi to Pennar) are sparse

Conclusion:

The data is highly skewed, with the Krishna Basin disproportionately overrepresented compared to other basins. This imbalance could distort hydrological analyses or resource allocation assessments, as smaller basins (e.g., Narmada, Indus) may be statistically overshadowed.

3. Water Storage Efficiency by State



Key Observations:

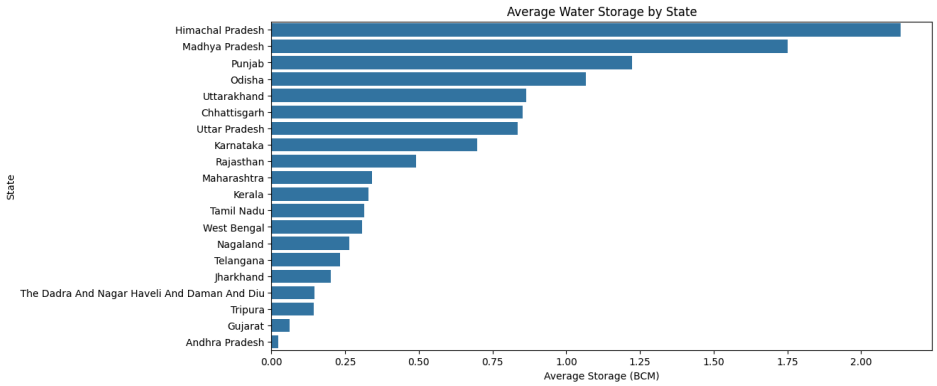
- Ukai Reservoir has the highest capacity significantly outpacing others.
- Kadana and Dharoi rank second and third.
- The remaining reservoirs (Panam to Sipu) show progressively lower capacities, all below.
- The top 3 reservoirs (Ukai, Kadana, Dharoi) collectively dominate the dataset, while the bottom 7 have modest capacities.

Conclusion:

The capacity distribution is highly unequal, with Ukai alone accounting for nearly as much as the next two largest reservoirs combined. This suggests:

- Regional dependency on a few key reservoirs (e.g., Ukai) for water storage.
- Potential vulnerability if top reservoirs face shortages or infrastructural issues.

4. Average Water Storage by State



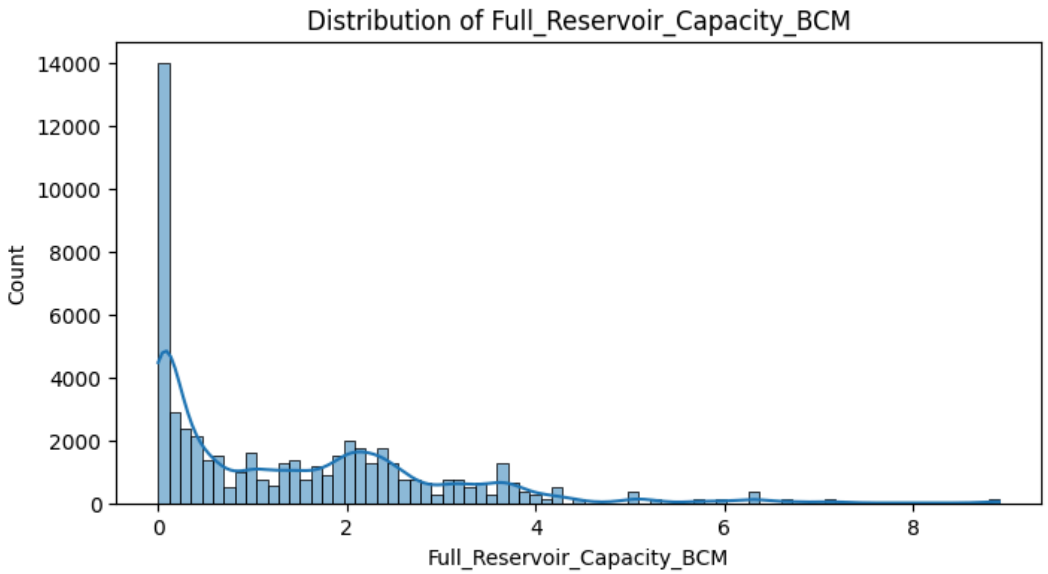
Key Observations:

- Himachal Pradesh has the highest average water storage
- Madhya Pradesh and Punjab follow with average storages above 1.0 BCM
- Odisha, Uttarakhand, and Chhattisgarh also report relatively high averages
- States like Andhra Pradesh, Gujarat, Tripura, and Dadra and Nagar Haveli and Daman and Diu have the lowest average storages, all well below 0.25 BCM.
- The majority of other states show moderate values between 0.3 to 0.8 BCM

Conclusion:

The graph highlights a clear imbalance in average water storage capacities among states. Himachal Pradesh stands out with the highest capacity, suggesting either large-scale water reservoirs or favorable geographic/hydrological conditions. In contrast, several states, particularly in western and northeastern India, have notably lower average storage. This disparity could have significant implications for water resource management, agricultural planning, and climate resilience strategies.

5. Distribution of Full_Reservoir_Capacity_BCM



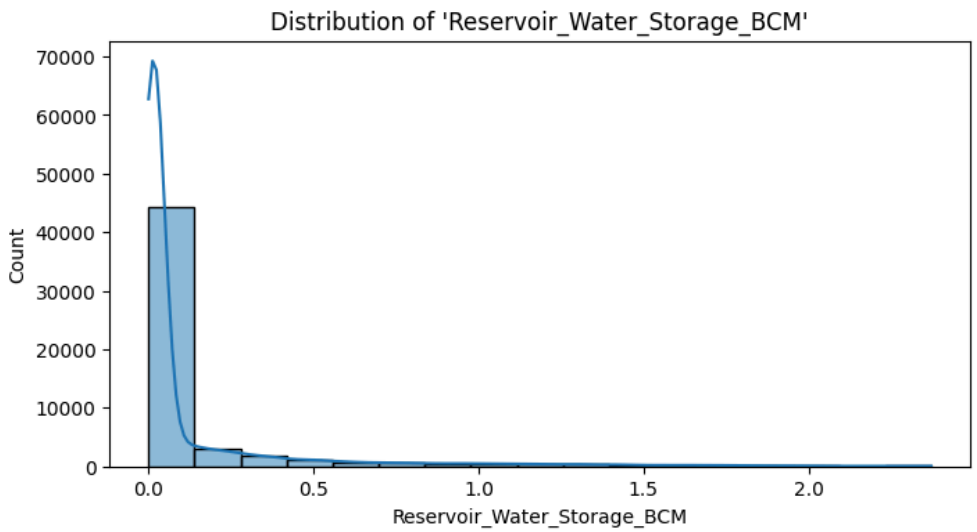
Key Observations:

- The distribution is heavily right-skewed, with a majority of reservoirs having a capacity of less than 1 BCM.
- The highest frequency is observed in the lowest bin (~0–0.5 BCM), where counts spike dramatically, with more than 14,000 entries.
- As reservoir capacity increases, the frequency gradually declines, forming a long tail extending up to around 8.5 BCM.
- Very few reservoirs have capacities beyond 5 BCM, indicating that large-capacity reservoirs are rare.

Conclusion:

The histogram shows a highly skewed distribution, suggesting that most reservoirs in the dataset are small in terms of capacity. A small number of large-capacity reservoirs contribute disproportionately to total water storage. This imbalance may affect regional water availability and infrastructure planning. Analytical models and resource policies should account for this skew to avoid overestimating storage capabilities based on a few large reservoirs.

6. Distribution of 'Reservoir_Water_Storage_BCM



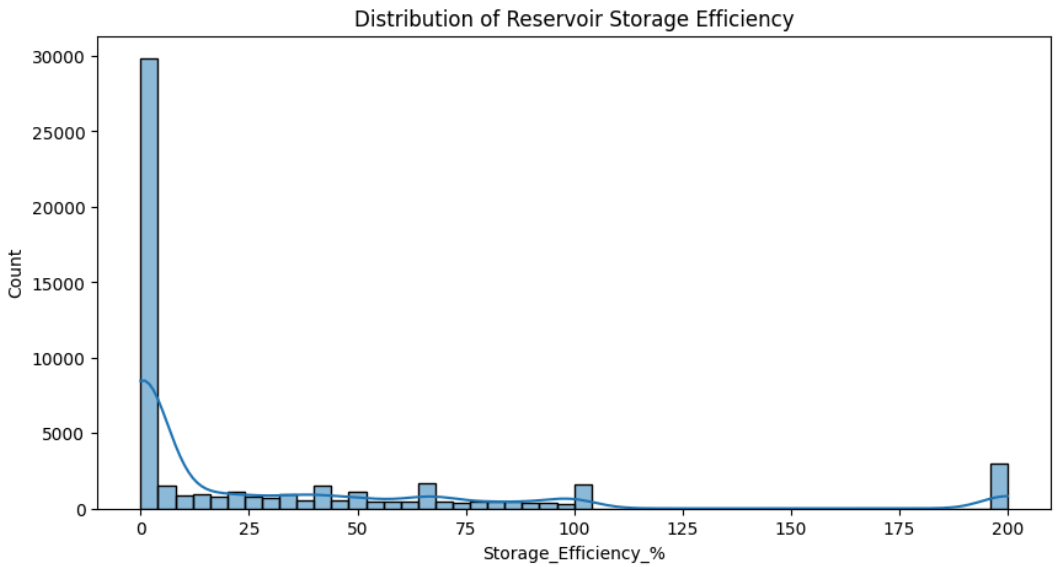
Key Observations:

- The distribution is extremely right-skewed, with an overwhelming number of records clustered below 0.1 BCM.
- The peak frequency exceeds 65,000 entries, showing that most recorded reservoirs have very small actual water storage volumes.
- Beyond 0.2 BCM, the count drops off rapidly, and values above 1.0 BCM are exceedingly rare.
- A long tail extends up to approximately 2.2 BCM, but these large storage volumes are outliers.

Conclusion:

This histogram reveals a stark concentration of reservoirs with very low actual water storage, indicating that most reservoirs are either small or underutilized at the time of data capture. The distribution is heavily skewed, suggesting potential issues such as seasonal depletion, under-capacity usage, or regional water stress. Any analysis or modeling using this variable must consider normalization or transformation techniques to handle the skew and avoid misleading results.

7. Distribution of Reservoir Storage Efficiency



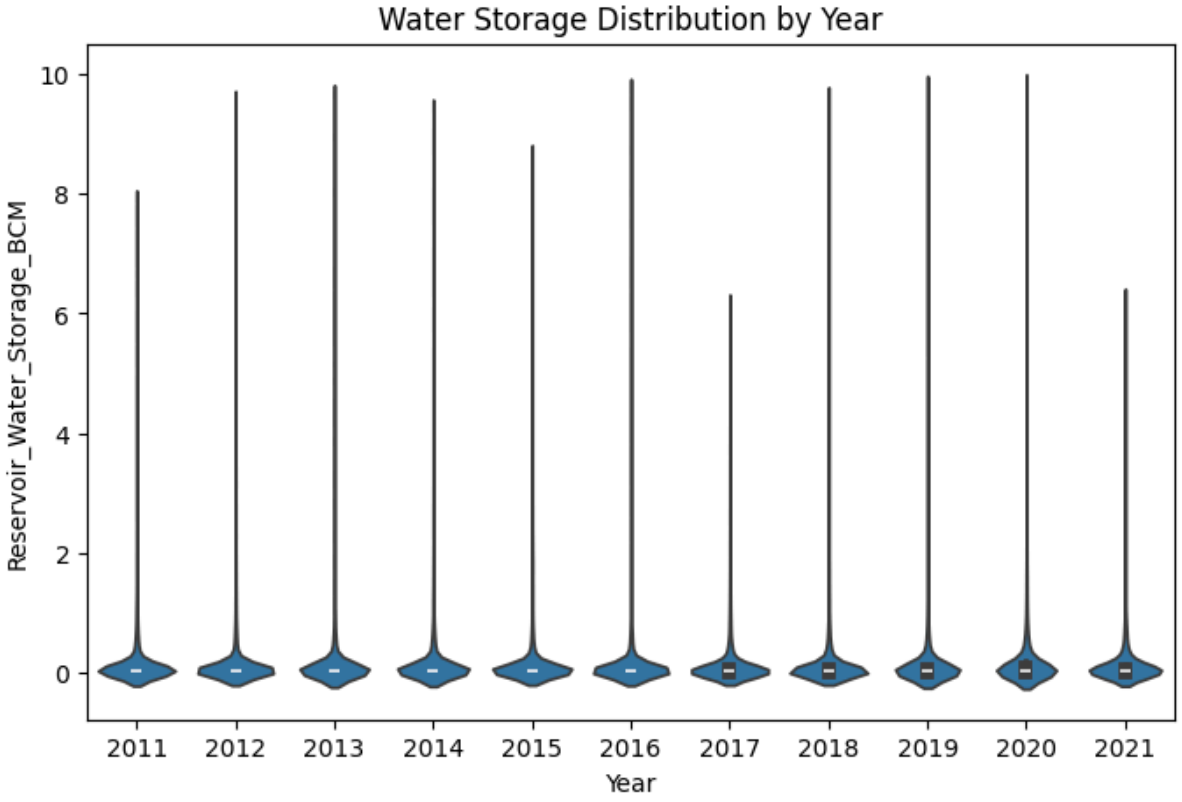
Key Observations:

- A significant number of reservoirs (~30,000 entries) have very low storage efficiency, falling in the 0–5% range, suggesting underutilization or inefficiencies.
- The distribution is highly skewed, with a steep decline in count as efficiency increases.
- There are multiple smaller peaks at regular intervals, including around 50%, 75%, and 100%, which may indicate specific operational or reporting thresholds.
- A noticeable spike exists at 200% efficiency, suggesting the presence of outliers or data anomalies, potentially due to incorrect inputs or unusual overflow conditions.

Conclusion:

The data reveals that most reservoirs operate at very low efficiency, which may point to systemic issues in water management or seasonal variations in reservoir usage. The presence of extreme values, including those above 100% efficiency, warrants further investigation or data cleaning, as they could distort analytical models. Overall, improving storage efficiency should be a key area of focus in reservoir planning and policy.

8. Water Storage Distribution by Year



Key Observations:

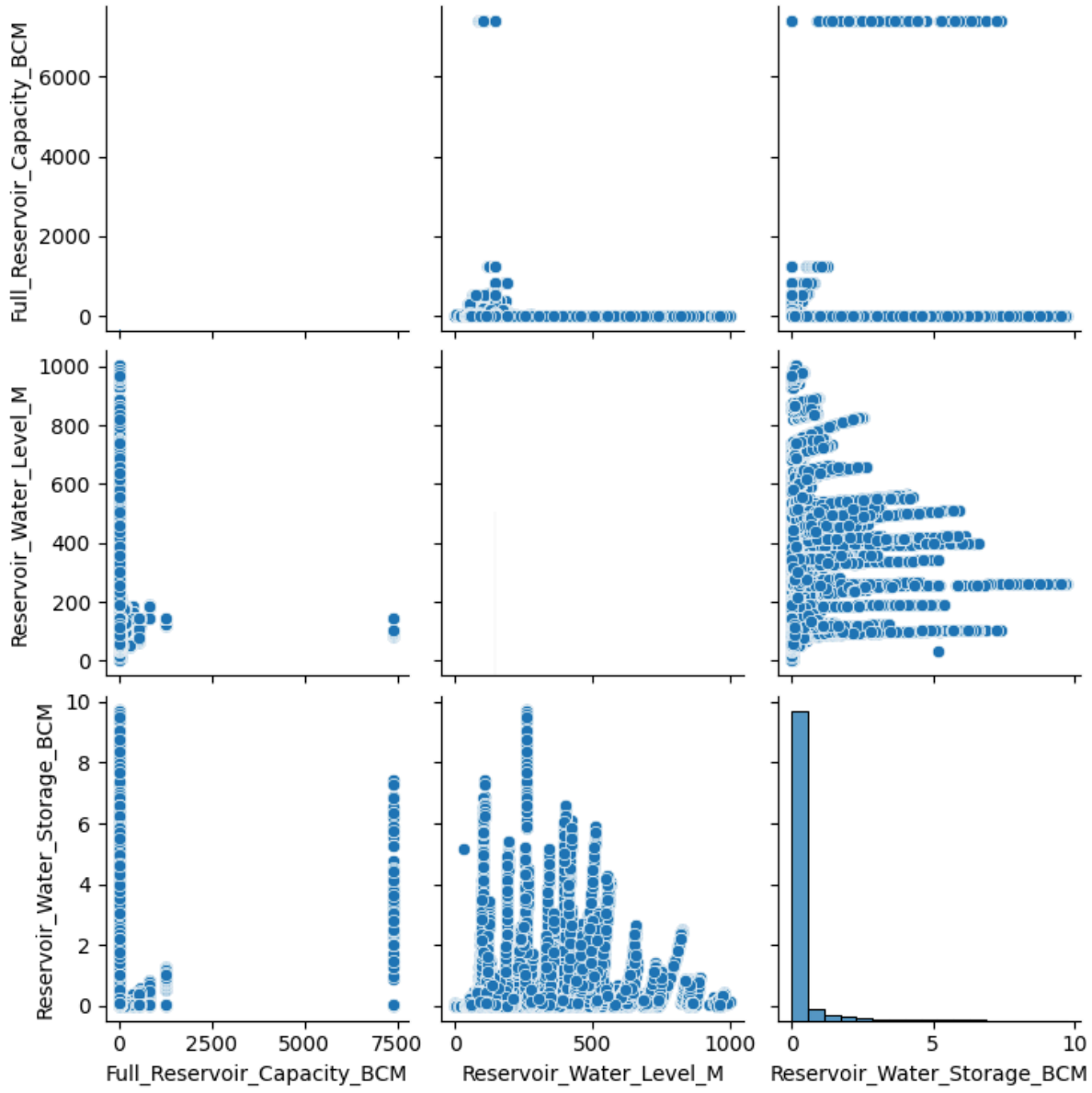
- Across all years from 2011 to 2021, the distribution of reservoir water storage is consistently right-skewed, with most data points clustered near 0 BCM.
- The median water storage remains consistently low throughout the years, indicating little variation in central tendency.
- However, each year shows the presence of long tails extending to high storage values (above 8–10 BCM), highlighting a few reservoirs with extremely large storage volumes.
- Notable dips in maximum storage spread are observed around 2017 and 2021, suggesting possible drought years or reduced storage reporting.

Conclusion:

The water storage distribution remains broadly consistent over the decade, with most reservoirs storing low volumes, while a few large reservoirs create long-tailed outliers. This imbalance in reservoir storage has persisted through the years, emphasizing the dominance of a small number of high-capacity reservoirs. Monitoring shifts in the upper range could be vital for understanding stress events such as floods or droughts, while the stable low median suggests continued underutilization or limited capacity across the majority of reservoirs.

7. Bivariate/Multivariate Analysis

1. Pair plot comparing “Full_Reservoir_Capacity_BCM”, “Reservoir_Water_Level_M”, and “Reservoir_Water_Storage_BCM”



Key Observations:

There is a visible positive correlation between:

- Reservoir_Water_Level_M and Reservoir_Water_Storage_BCM – as water level increases, water storage also increases.

- However, the relationship is nonlinear and scattered, indicating that depth alone doesn't fully determine volume stored (possibly due to varying shapes or surface areas of reservoirs).

The relationship between Full_Reservoir_Capacity_BCM and both Water Level and Water Storage appears highly skewed and clustered near the origin, suggesting that:

- Most reservoirs have low capacity.
- A few outliers with extremely high capacities heavily influence the range.

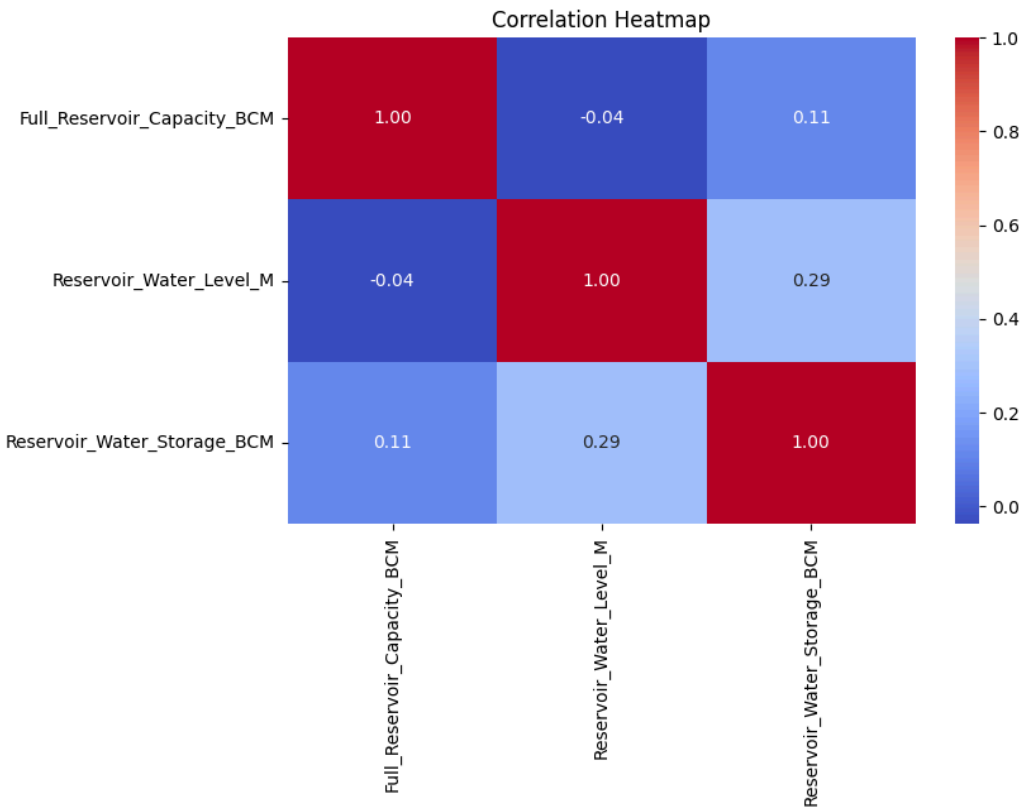
The diagonal plots reinforce earlier findings:

- Most values for Reservoir_Water_Storage_BCM and Full_Reservoir_Capacity_BCM are concentrated near zero.
- The Reservoir_Water_Level_M has a wider spread but still shows clustering at lower ranges.
- Presence of vertical and horizontal lines in scatter plots indicates repeated values, possibly due to fixed measurement thresholds or standard designs.

Conclusion:

The pairwise relationships show that while water level positively influences water storage, the variation in reservoir design (capacity, geometry) leads to a nonlinear and diffuse correlation. The high concentration of values near zero across all variables further confirms that small reservoirs dominate the dataset. Any predictive modeling (e.g., estimating storage from level or capacity) will need to account for outliers and the heterogeneous nature of reservoir infrastructure.

2. Correlation Heatmap



Key Observations:

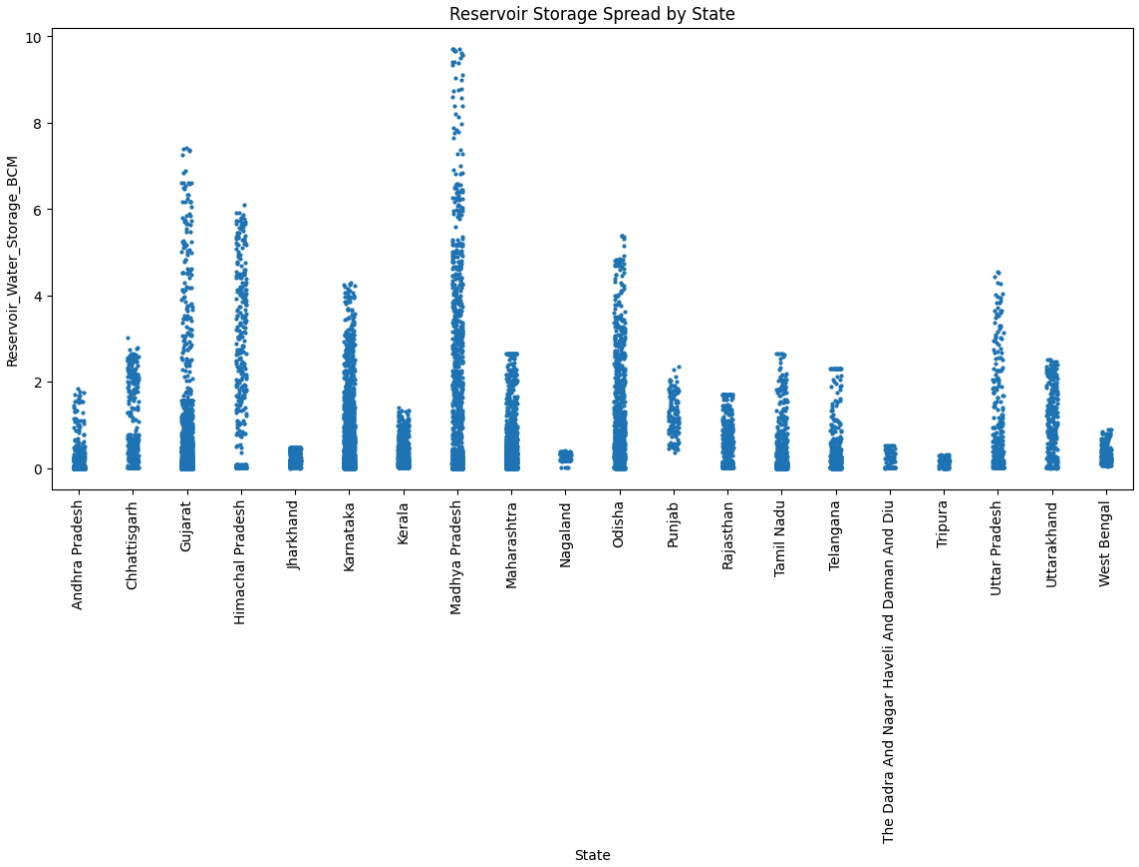
Reservoir_Water_Storage_BCM has:

- A moderate positive correlation (0.29) with Reservoir_Water_Level_M, indicating that as water level increases, water storage tends to increase, though the relationship is not strong.
- A weak positive correlation (0.11) with Full_Reservoir_Capacity_BCM, implying that higher capacity doesn't strongly guarantee higher storage at a given time.
- Full_Reservoir_Capacity_BCM and Reservoir_Water_Level_M show a very weak negative correlation (-0.04), essentially indicating no meaningful relationship between a reservoir's full capacity and its current water level.

Conclusion:

The heatmap highlights that reservoir water storage is influenced more by water level than by full capacity, but neither variable shows a strong linear relationship with it. This suggests that other factors such as rainfall, catchment area, or operational policies may significantly influence water storage. The low correlations emphasize the complexity and variability of reservoir behavior, and caution should be used when attempting to predict storage based solely on level or capacity.

3. Reservoir Storage Spread by State



Key Observations:

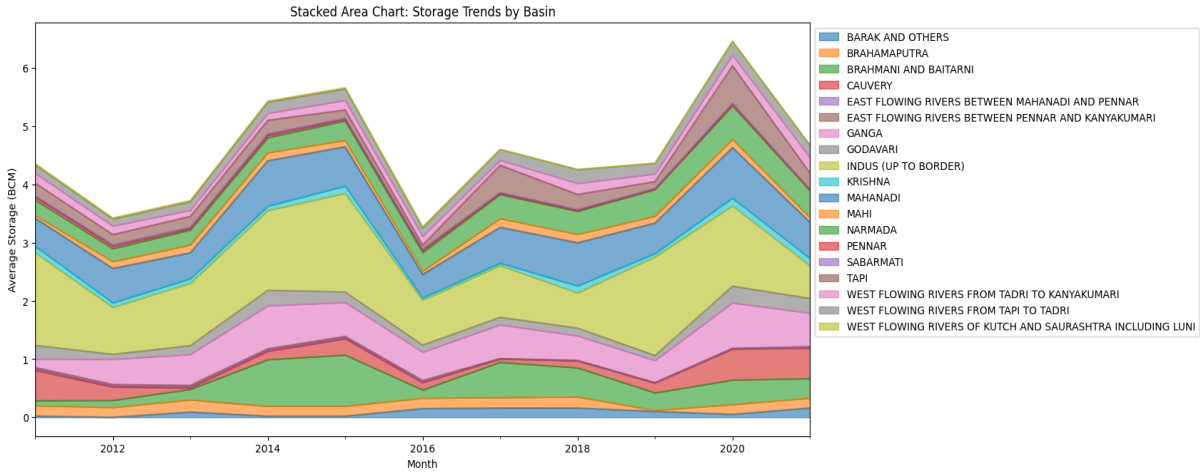
1. States with the Highest Storage Volumes:
    - Madhya Pradesh and Maharashtra show the widest spread and highest values of water storage (up to ~10 BCM).
    - Gujarat and Odisha also demonstrate relatively high storage values and wider data point dispersion.
2. Moderate Storage States:
    - Andhra Pradesh, Chhattisgarh, Tamil Nadu, Rajasthan, Karnataka, Telangana, and Uttar Pradesh have mid-range storage levels (mostly between 0 and 4 BCM), indicating a balance between demand and supply or moderately sized reservoirs.
3. Low Storage/Small Reservoir States:
    - Nagaland, Tripura, Jharkhand, Uttarakhand, Himachal Pradesh, and The Dadra and Nagar Haveli, and Daman and Diu show lower storage volumes, likely due to geographic limitations (hills, small river basins) or fewer reservoirs.

Conclusion:

- Madhya Pradesh and Maharashtra play a critical role in India's reservoir water storage infrastructure, likely due to their large land area.
- Eastern and northeastern states (like Tripura and Nagaland) show minimal storage, highlighting the regional disparity in water infrastructure.
- The distribution is highly skewed, suggesting that national water policy must consider regional imbalances and boost storage capacities.
- States like Gujarat and Odisha show efficient use of water storage despite not having the largest landmass, which could be worth studying for best practices.

8. Trend/Time Analysis

1. Storage Trends by Basin



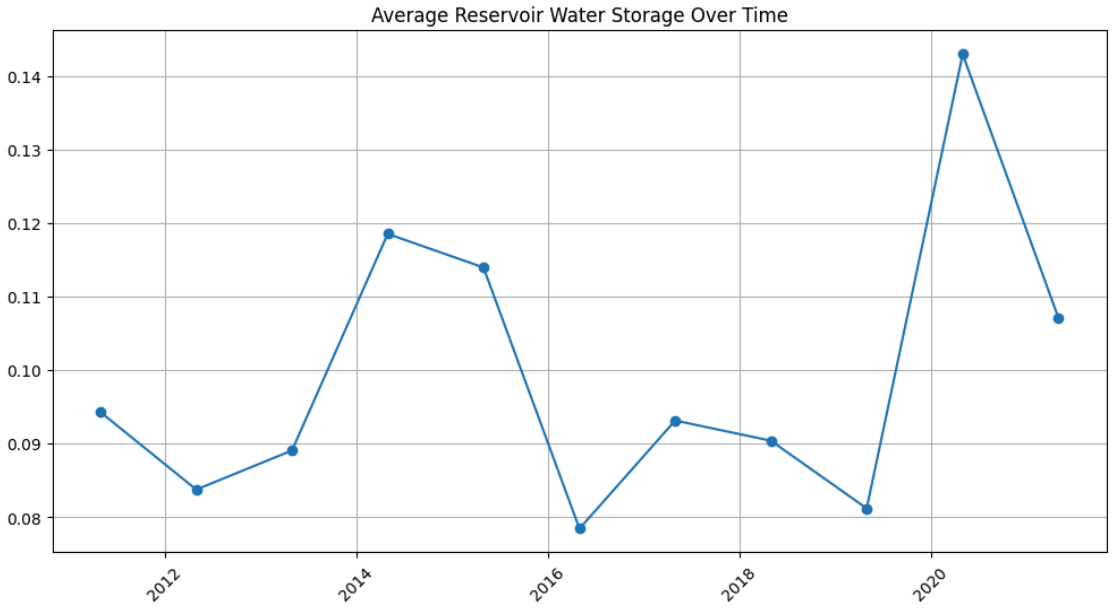
Key Observations:

- Godavari, Krishna, Mahanadi, and West Flowing Rivers of Kutch and Saurashtra consistently contribute the largest share of water storage.
- Storage peaks around 2014, 2016, and 2020 likely correspond to stronger monsoon years, while dips in 2012 and 2016 may indicate dry periods.
- Smaller basins like Barak, Brahmaputra, Pennar, and Sabarmati show minimal contribution throughout the period.
- Storage trends show cyclical variation, with no strong long-term increase or decrease, pointing to high dependence on annual rainfall and climatic events.

Conclusion:

The chart reveals that a few major basins dominate national storage patterns, while most others play a limited role. Fluctuations are largely climate-driven, emphasizing the need for adaptive water management that accounts for variability across regions and years. Overall, rainfall and basin characteristics appear more influential than reservoir capacity alone.

2. Average reservoir water storage over time



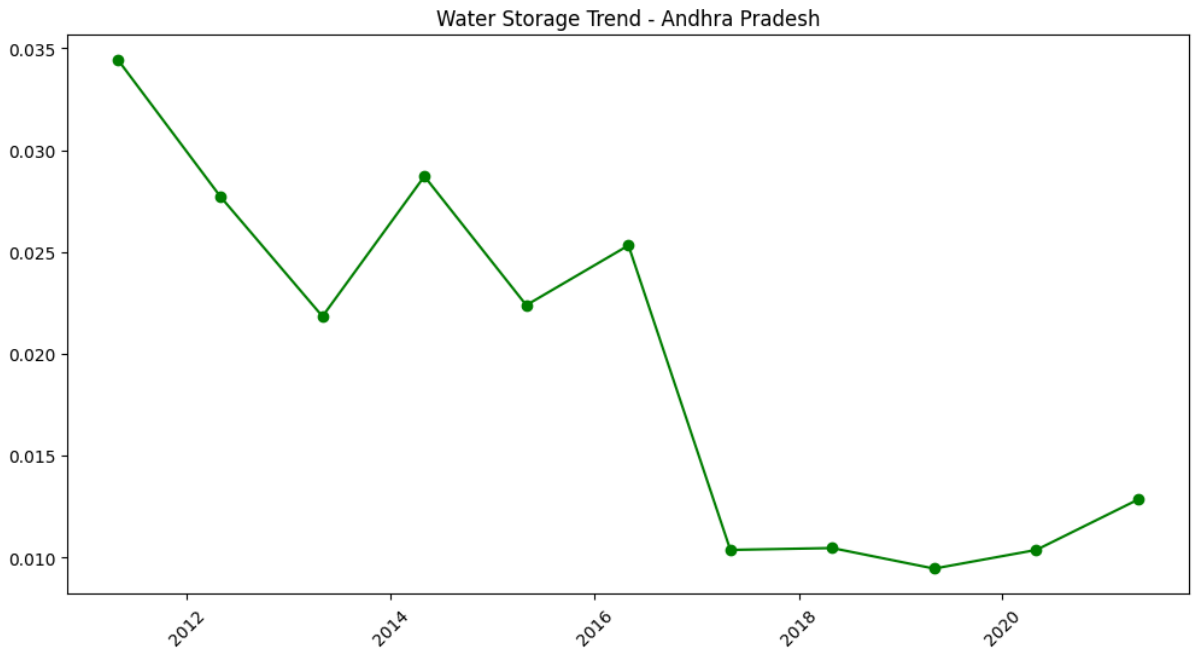
Key Observations:

- The average water storage shows noticeable year-to-year variability.
- Highest storage occurred in 2020, followed by a sharp decline in 2021.
- Lows were seen in 2012, 2016, and 2019, suggesting years of lower inflow or drought-like conditions.
- Upticks in 2015 and 2017 indicate partial recoveries, but the pattern remains inconsistent.

Conclusion:

Average reservoir storage over time reflects strong inter-annual fluctuations, likely driven by monsoon performance and climatic factors. The peak in 2020 stands out, but overall the data shows no consistent upward or downward trend, highlighting the importance of seasonal planning and water conservation policies.

3. Average reservoir water storage over time in Andhra Pradesh



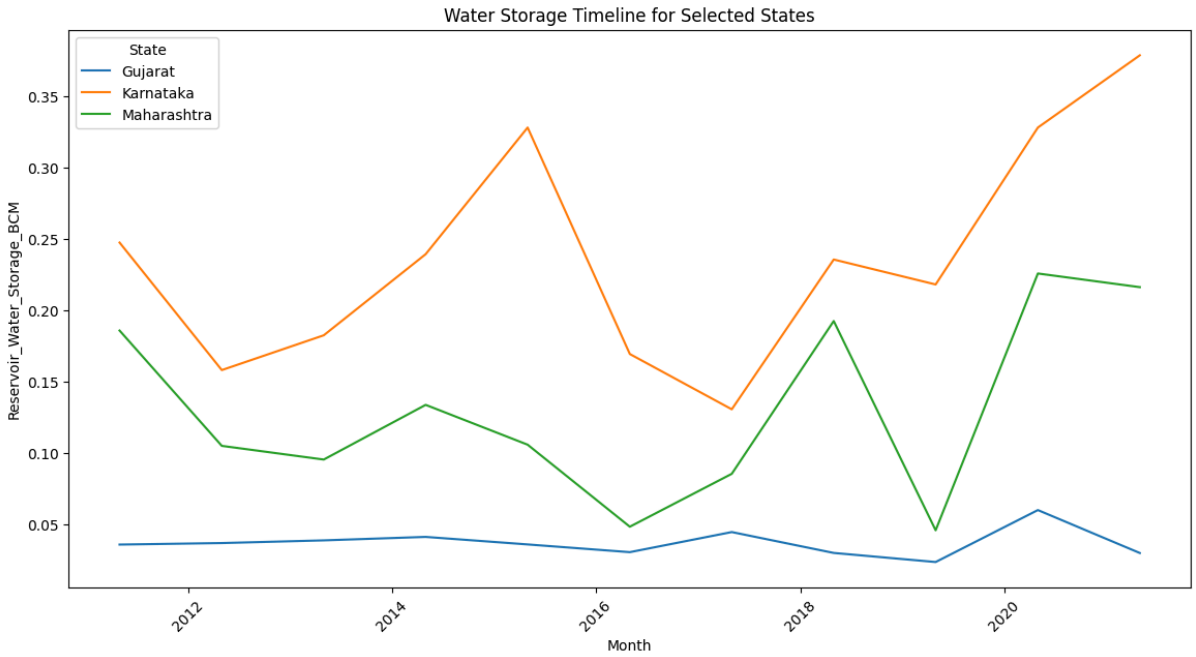
Key Observations:

- Highest storage was recorded in 2011, followed by a steady decline until 2013.
- A brief recovery occurred in 2014 and 2016, but overall the trend continues downward until 2017.
- From 2017 onward, water storage remained consistently low, with a slight upward shift after 2019.
- The lowest point occurred around 2019, indicating likely water stress or drought conditions.

Conclusion:

Andhra Pradesh's water storage shows a clear declining trend over the decade, suggesting growing water scarcity or reduced inflows. While there are short recoveries, post-2016 storage levels remain significantly below earlier years. This trend highlights the urgent need for sustainable water management, including better storage infrastructure, rainwater harvesting, and efficient usage policies.

4. Average reservoir water storage timeline for different States



Key Observations:

- Karnataka consistently has the highest reservoir storage, showing a strong upward trend from 2017 to 2021, peaking in the last year.
- Maharashtra shows high variability, with major drops in 2016 and 2019, followed by a sharp rise in 2020.
- Gujarat maintains the lowest storage levels, with relatively stable but low values and minor fluctuations over the years.

Conclusion:

Among the three states, Karnataka shows the most consistent and improving storage trend, possibly due to better inflow, rainfall, or infrastructure. Maharashtra's pattern is more erratic, indicating a need for improved water management. Gujarat's persistently low storage levels suggest limited reservoir capacity or regional water stress, requiring strategic attention.

9. References

Dataset: https://ndap.niti.gov.in/dataset/7062?filter_id=4509

Github: https://github.com/gautam0222/WaterReservoirEDA