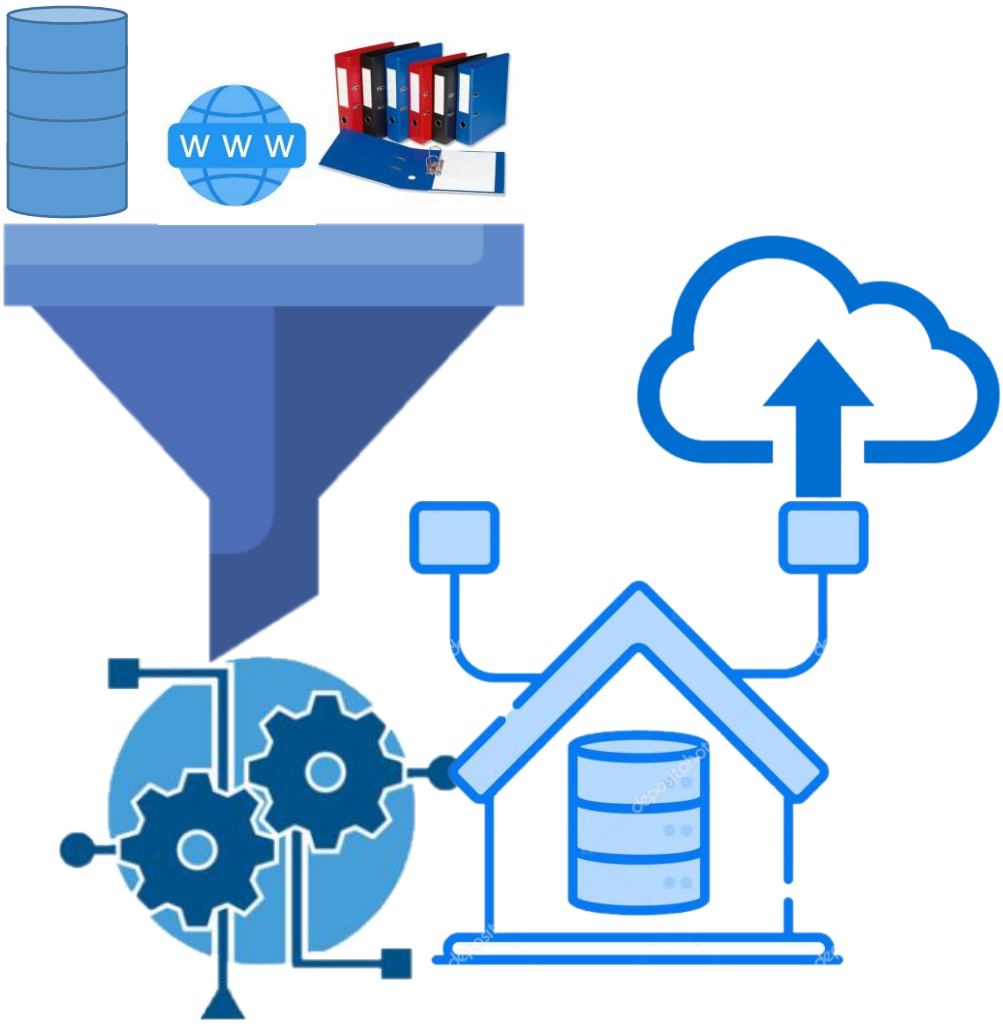
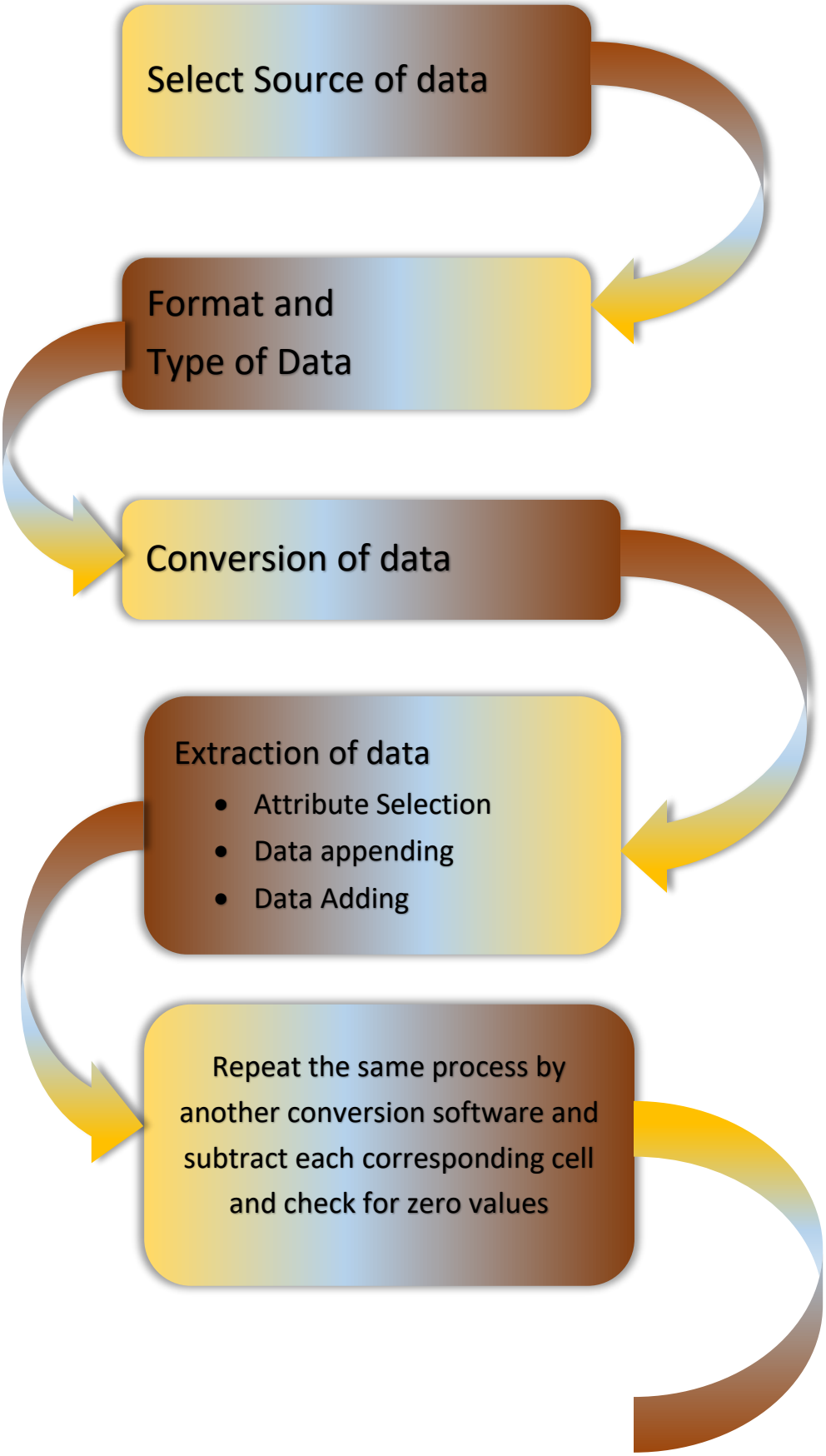
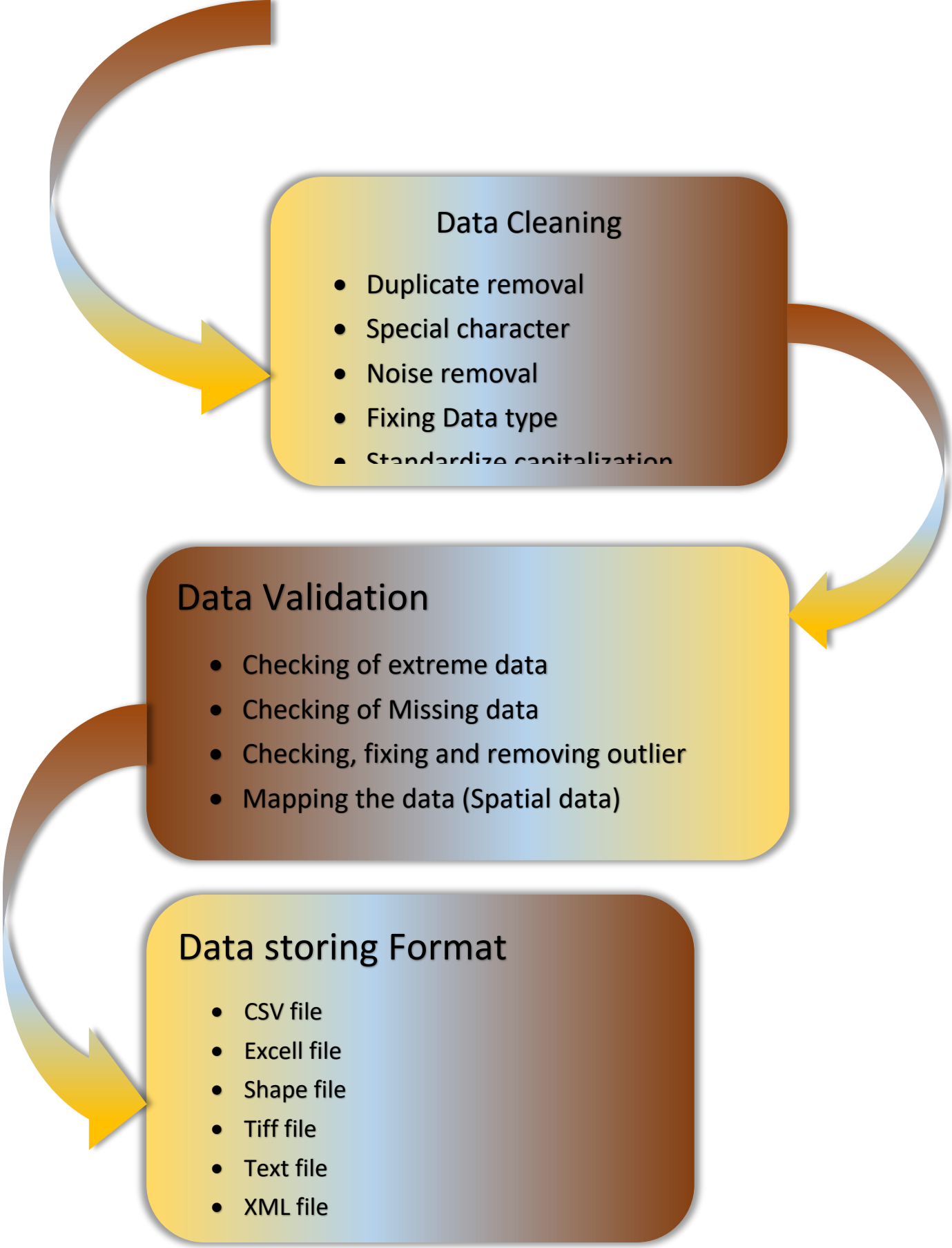
ETL Pipeline for CPCB Data



ETL Flow diagram

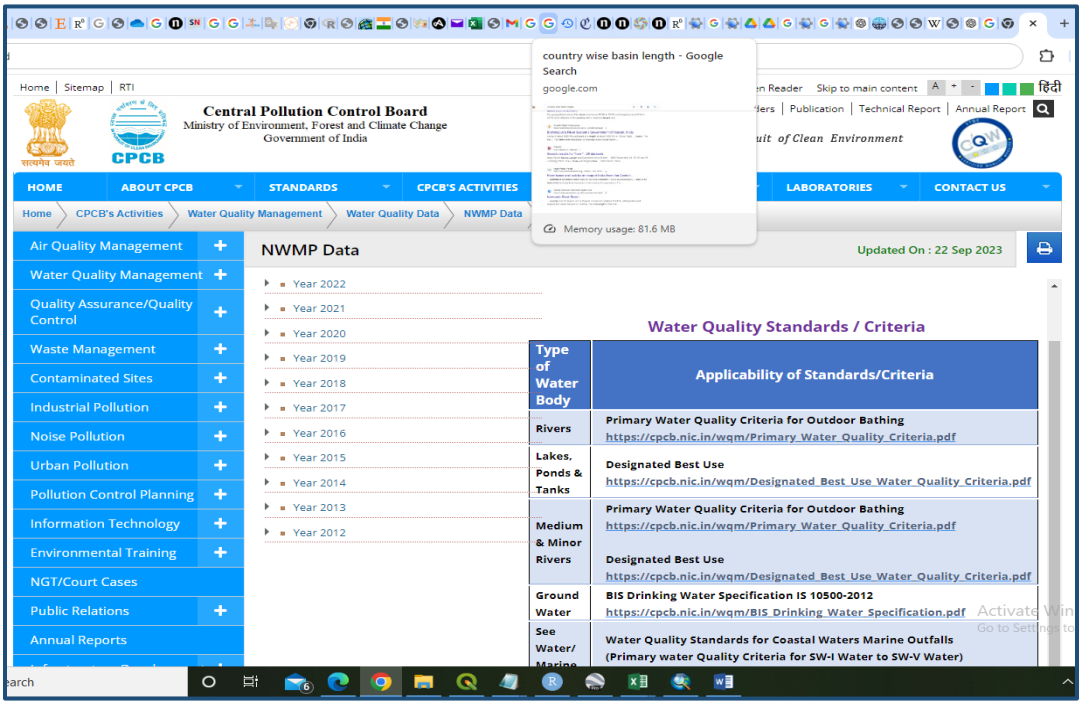



Data Cleaning
Duplicate removal
Special character
Noise removal
Fixing Data type
Data Validation
Checking of extreme data
Checking of Missing data
Checking, fixing and removing outlier
Mapping the data (Spatial data)
Data storing Format
CSV file
Excell file
Shape file
Tiff file
Text file
XML file

Open central pollution control board Website specifically NWMP DATA

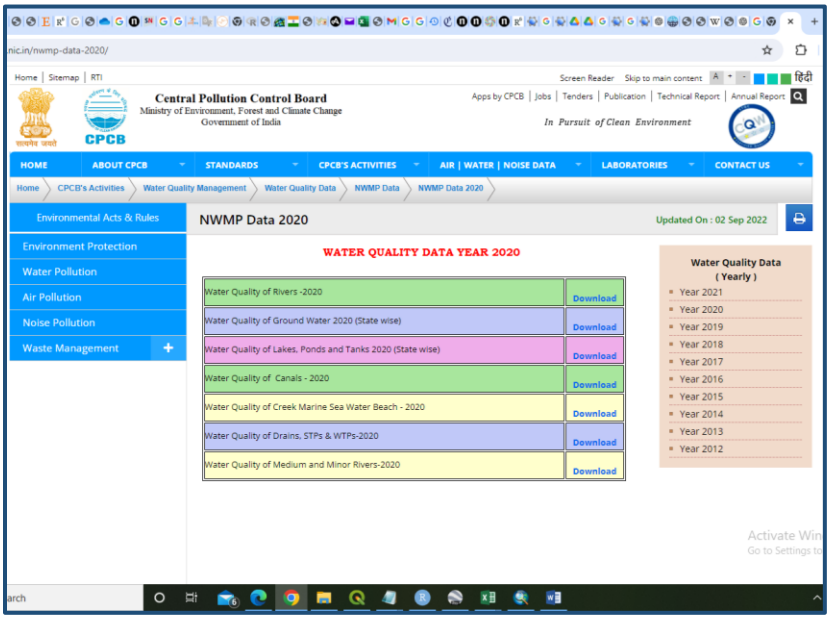
https://cpcb.nic.in/nwmp-data/

Select specific year data i.e Year 2021



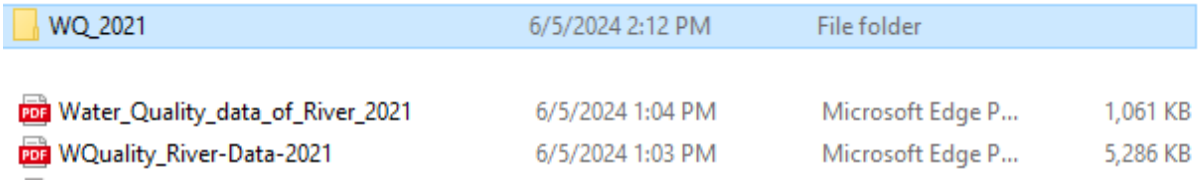
Download Water Quality of River-2021

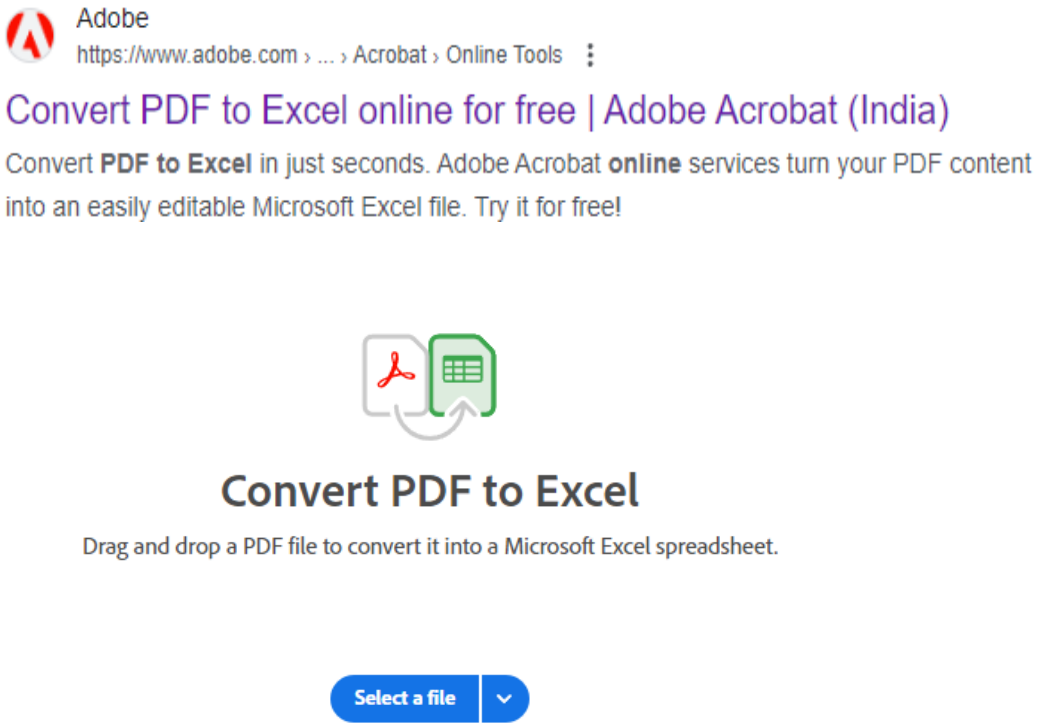
Download Water Quality of Medium and Minor Rivers-2021



Select Source of data

Download both the fila and save in folder name WQ_2020

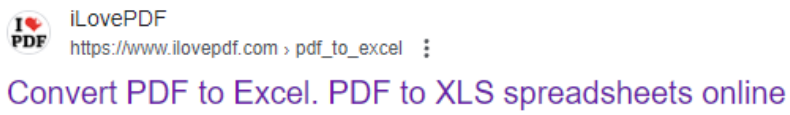




Select file from your folder WQ_2021

Down load the converted files in to your WQ_2021 folder

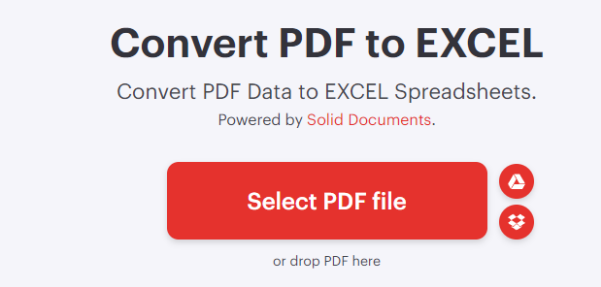
This software is paid



Format and Type of Data
(pdf file and text and Numeric data)

Conversion of data

(Select two different pdf to excel converter software)

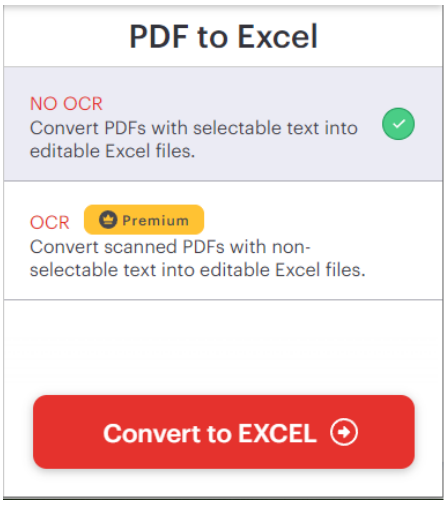


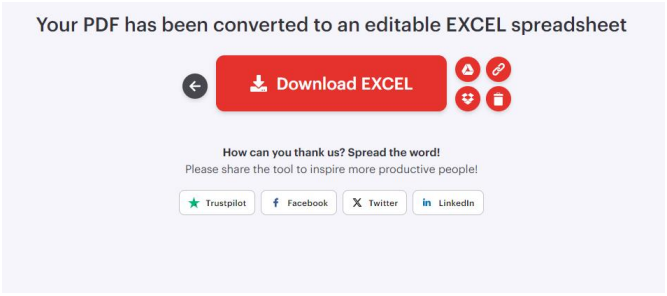
Select or drop Pdf file from your Directory and upload.

Click convert to excel button.

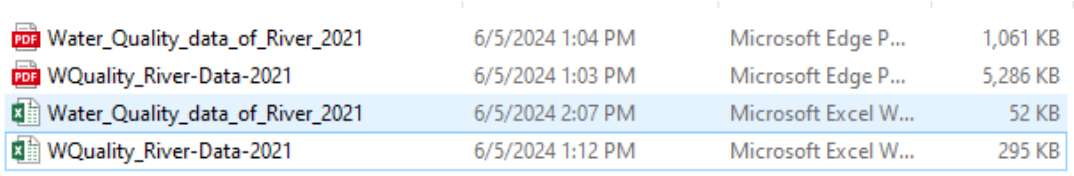
After few second or minute it convert to excel.

Now down load it in your desired directory.





There are two excel files



Open the excel file.

Extraction of data

- Attribute Selection

Water Quality of Medium & Minor Rivers under NWMP-2021																	
STN Code	Name of Monitoring Location	State Name	Temperature °C		Dissolved Oxygen (mg/L)		pH		Conductivity (µmho/cm)		BOD (mg/L)		Nitrate N + Nitrite N(mg/L)		Fecal Coliform (MPN/100ml)		Total Coliform (MPN/100ml)	
			Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
1448	RIVER NAGAVALI AT THOTAPALLI REGULATOR, VIZIANAGARAM	ANDHRA PRADESH	23.0	32.0	5.9	8.2	7.0	8.2	242	400	1.4	2.0	0.72	1.66	7	15	93	150
4351	RIVER NAGAVALI IMMEDIATE BORDER BETWEEN ANDHRA PRADESH & ORISSA STATES, NEAR SIVALAYAM, KUNERU (V), KOMARADA (M)	ANDHRA PRADESH	24.0	32.0	5.2	8.3	6.7	8.5	240	612	1.2	2.5	0.92	3.02	11	21	150	210
4346	RIVER NAGAVALI NEAR NH- 16 BRIDGE, UPSTREAM (U/S) OF SRIKAKULAM TOWN, BEFORE CONFLUENCE OF TOWN SEWAGE	ANDHRA PRADESH	24.0	32.0	6.4	8.4	7.1	8.8	243	400	1.2	2.0	0.54	3.60	7	15	120	120
4347	RIVER NAGAVALI NEAR WATER PUMP HOUSE (KILLIPALEM) , DOWN STREAM (D/S) OF SRIKAKULAM TOWN, AFTER CONFLUENCE OF TOWN SEWAGE	ANDHRA PRADESH	24.0	32.0	6.0	7.5	7.1	8.4	330	648	1.4	2.3	0.33	4.02	11	15	150	210
4348	RIVER VAMSADHARA, IMMEDIATE BORDER BETWEEN ANDHRA PRADESH & ORISSA STATES, NEAR LALITHAMBA TEMPLE, BATTILI (V) BHAMINI (M)	ANDHRA PRADESH	22.0	28.0	5.0	8.3	6.9	8.5	221	410	1.5	2.2	0.42	2.46	9	21	120	210
2352	RIVER VAMSADHARA, KALINGAPATNAM, VIZIANAGARAM	ANDHRA PRADESH	24.0	32.0	5.8	8.8	7.0	8.3	228	420	1.5	2.2	0.42	3.05	7	15	120	150
1393	RIVER DAMANGANGA AT D/S OF MADHUBAN, DAMAN	DAMAN AND DIU, DADRA AND NAGAR HAVELI	26.2	30.0	6.6	7.9	7.6	8.5	104	251	1.0	1.0	0.30	0.30	5	9	24	39
2461	RIVER DAMANGANGA AT DAMAN JETTY, MOTI DAMAN	DAMAN AND DIU, DADRA AND NAGAR HAVELI	27.1	29.0	2.7	6.7	6.7	8.3	19214	25108	1.0	7.1	0.50	0.90	21	53	119	346
2460	RIVER DAMANGANGA AT DISCHARGE POINT OF DISTILLERY, DAMAN	DAMAN AND DIU, DADRA AND NAGAR HAVELI	27.1	30.0	4.8	6.1	6.7	8.0	15824	17943	2.6	6.1	0.30	0.50	36	110	297	427
2463	RIVER DAMANGANGA AT LAVACHA TEMPLE, SILVASSA	DAMAN AND DIU, DADRA AND NAGAR HAVELI	26.8	30.0	6.4	7.7	7.5	8.6	131	406	1.0	1.0	0.30	0.30	35	41	74	91
2465	RIVER DAMANGANGA AT NAROLI BRIDGE, SILVASSA	DAMAN AND DIU, DADRA AND	26.4	30.5	6.6	7.5	7.4	8.9	112	277	1.0	1.0	0.30	0.50	12	18	31	38

Table 1

Ready

Delete column “Name of Monitoring Location” and “State Name”

Water Quality of Medium & Minor Rivers under NWMP-2021																
STN Code	Temperature °C		Dissolved Oxygen (mg/L)		pH		Conductivity (µmho/cm)		BOD (mg/L)		Nitrate N + Nitrite N(mg/L)		Fecal Coliform (MPN/100ml)		Total Coliform (MPN/100ml)	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
1448	23.0	32.0	5.9	8.2	7.0	8.2	242	400	1.4	2.0	0.72	1.66	7	15	93	150
4351	24.0	32.0	5.2	8.3	6.7	8.5	240	612	1.2	2.5	0.92	3.02	11	21	150	210

Replace top three rows by following rows

st_Code	Tmp_Min_21	Tmp_Max_21	Do_Min_21	Do_Max_21	pH_Min_21	pH_Max_21	Con_Min_21	Con_Max_21	Bod_Min_21	Bod_Max_21	Nit_Min_21	Nit_Max_21	Fa_Co_Min_21	Fa_Co_Max_21	To_Co_Min_21	To_Co_Max_21

Here Tmp_Min_21: Minimum Temperature of Year 2021

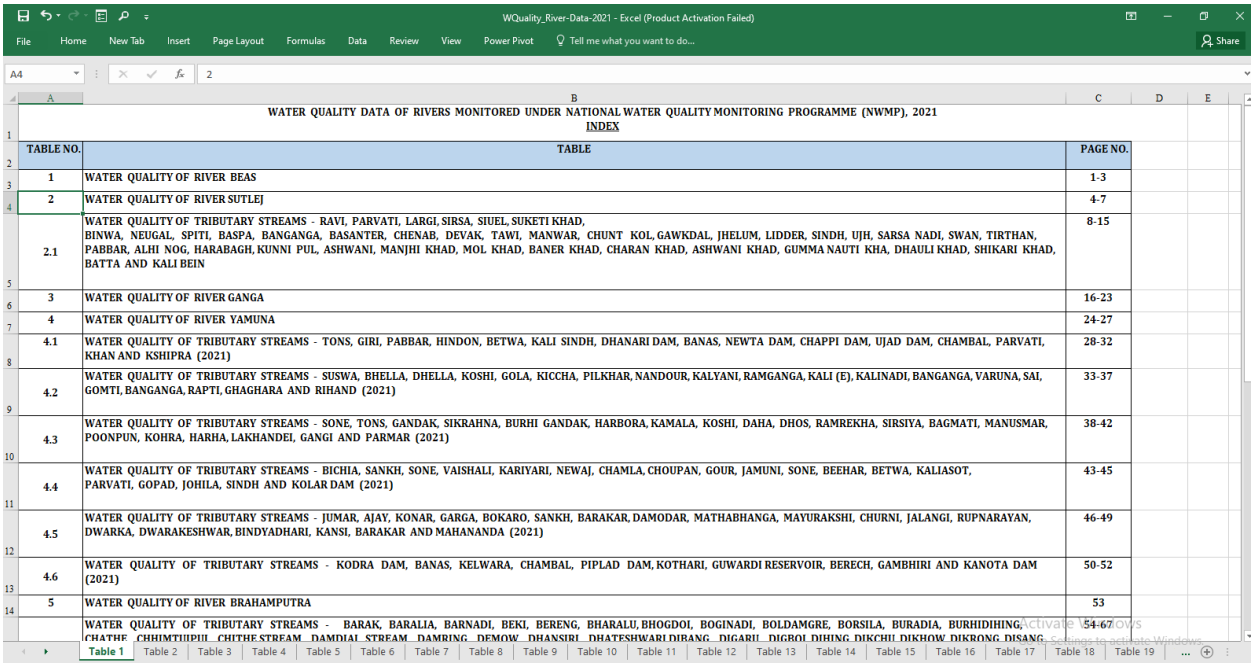
Similarly make appropriate column headings for other parameters

Now the top row looking like following excel sheet

st_Code	Tmp_Min_20	Tmp_Max_21	Do_Min_21	Do_Max_21	pH_Min_21	pH_Max_21	Con_Min_21	Con_Max_21	Bod_Min_21	Bod_Max_21	Nit_Min_21	Nit_Max_21	Fa_Co_Min_21	Fa_Co_Max_21	To_Co_Min_21	To_Co_Max_21
1448	23.0	32.0	5.9	8.2	7.0	8.2	242	400	1.4	2.0	0.72	1.66	7	15	93	150
4351	24.0	32.0	5.2	8.3	6.7	8.5	240	612	1.2	2.5	0.92	3.02	11	21	150	210
4346	24.0	32.0	6.4	8.4	7.1	8.8	243	400	1.2	2.0	0.54	3.60	7	15	120	120
4347	24.0	32.0	6.0	7.5	7.1	8.4	330	648	1.4	2.3	0.33	4.02	11	15	150	210
4348	22.0	28.0	5.0	8.3	6.9	8.5	221	410	1.5	2.2	0.42	2.46	9	21	120	210
2352	24.0	32.0	5.8	8.8	7.0	8.3	228	420	1.5	2.2	0.42	3.05	7	15	120	150
1393	26.2	30.0	6.6	7.9	7.6	8.5	104	251	1.0	1.0	0.30	0.30	5	9	24	39

Now save the file as "**Water_Quality_Minor_River_2021**"

Now for another file which have water quality data of rivers. It contains data in many tables.



First table shows the index document, it need not to change anything.

Table 2 sheet has the water quality data of river beas

TABLE 1 -: WATER QUALITY DATA OF RIVER BEAS (2021)

STATION CODE	NAME OF MONITORING LOCATION	STATE NAME	TEMPERATURE (°C)		DISSOLVED OXYGEN (mg/L)		pH		CONDUCTIVITY (µmhos/cm)		BIO- CHEMICAL OXYGEN DEMAND (mg/L)		NITRATE (mg/L)		FECAL COLIFORM (MPN/100ML)		TOTAL COLIFORM (MPN/100ML)		FECAL STREPTOCOCCI (MPN/100ML)	
			MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX
PRIMARY WATER QUALITY CRITERIA NOTIFIED UNDER E(P) RULES, 1986					> 5.0		6.5-8.5				< 3.0				< 2500 MPN/100ML				< 500 MPN/100ML	
1001	RIVER BEAS AT U/S MANALI	HIMACHAL PRADESH	4.0	14.0	7.9	9.7	7.2	8.0	80	270	1.0	1.0	0.30	0.72	2	110	25	540	2	2

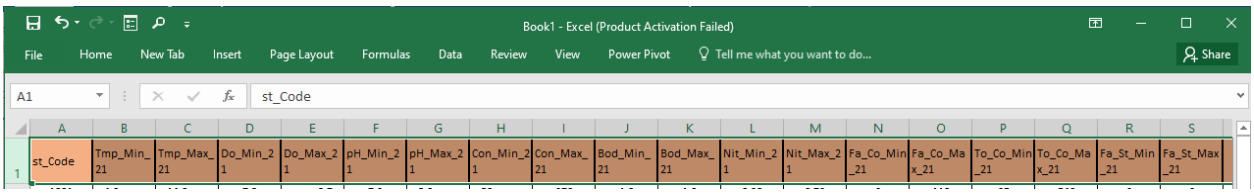
Delete column "Name of Monitoring Location" and "State Name" and replace top 5 rows by following row

Extraction of data

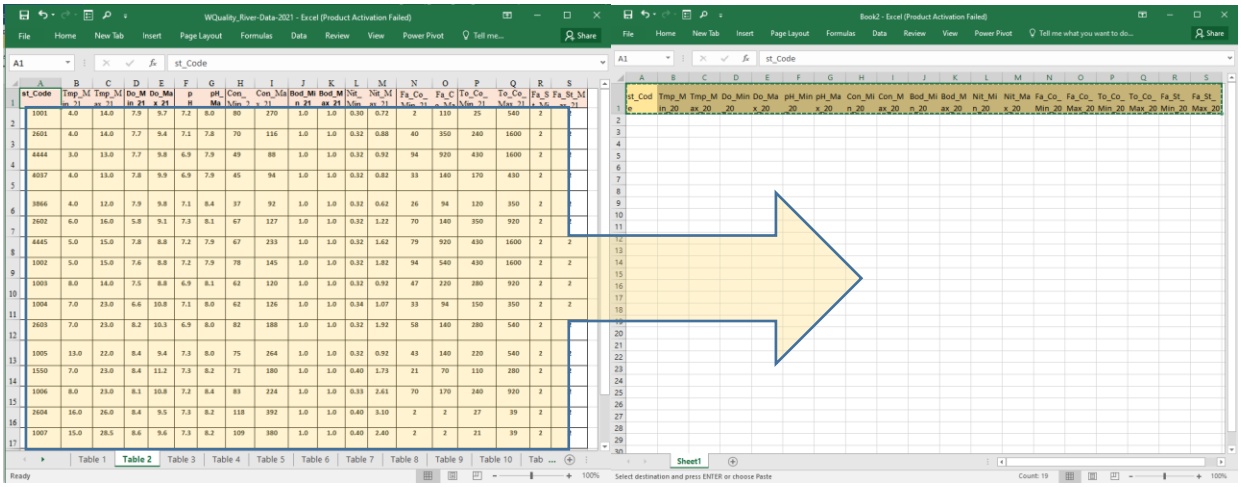
- Data appending
- Data Adding

st_Code	Tmp_Min_21	Tmp_Max_21	Do_Min_21	Do_Max_21	pH_Min_21	pH_Max_21	Con_Min_21	Con_Max_21	Bod_Min_21	Bod_Max_21	Nit_Min_21	Nit_Max_21	Fa_Co_Min_21	Fa_Co_Max_21	To_Co_Min_21	To_Co_Max_21	Fa_St_Min_21	Fa_St_Max_21

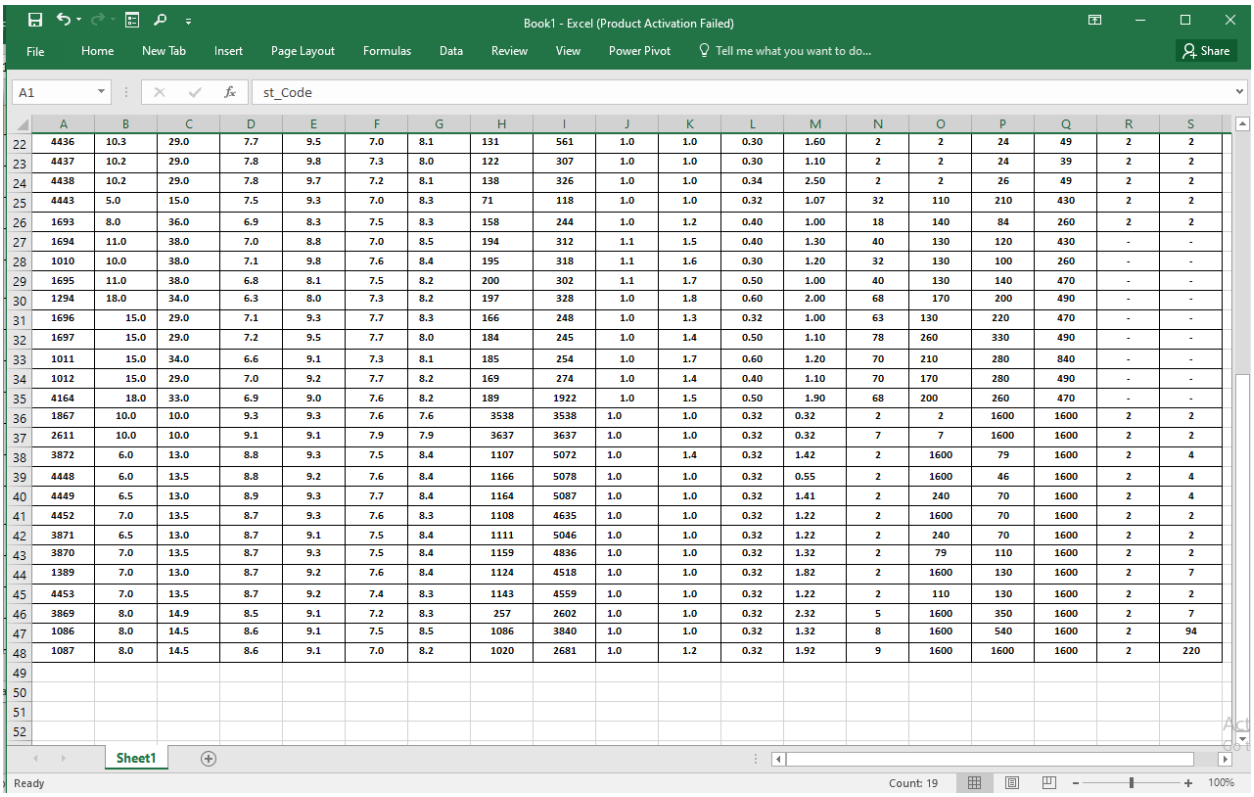
Open a new excel file and keep the top row as above mentioned



Copy the data from table 2 to new excel sheet



Similarly for table 3, delete two column and copy the data portion and append in new sheet where previously tables data was copied.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
22	4436	10.3	29.0	7.7	9.5	7.0	8.1	131	561	1.0	1.0	0.30	1.60	2	2	24	49	2	2
23	4437	10.2	29.0	7.8	9.8	7.3	8.0	122	307	1.0	1.0	0.30	1.10	2	2	24	39	2	2
24	4438	10.2	29.0	7.8	9.7	7.2	8.1	138	326	1.0	1.0	0.34	2.50	2	2	26	49	2	2
25	4443	5.0	15.0	7.5	9.3	7.0	8.3	71	118	1.0	1.0	0.32	1.07	32	110	210	430	2	2
26	1693	8.0	36.0	6.9	8.3	7.5	8.3	158	244	1.0	1.2	0.40	1.00	18	140	84	260	2	2
27	1694	11.0	38.0	7.0	8.8	7.0	8.5	194	312	1.1	1.5	0.40	1.30	40	130	120	430	-	-
28	1010	10.0	38.0	7.1	9.8	7.6	8.4	195	318	1.1	1.6	0.30	1.20	32	130	100	260	-	-
29	1695	11.0	38.0	6.8	8.1	7.5	8.2	200	302	1.1	1.7	0.50	1.00	40	130	140	470	-	-
30	1294	18.0	34.0	6.3	8.0	7.3	8.2	197	328	1.0	1.8	0.60	2.00	68	170	200	490	-	-
31	1696	15.0	29.0	7.1	9.3	7.7	8.3	166	248	1.0	1.3	0.32	1.00	63	130	220	470	-	-
32	1697	15.0	29.0	7.2	9.5	7.7	8.0	184	245	1.0	1.4	0.50	1.10	78	260	330	490	-	-
33	1011	15.0	34.0	6.6	9.1	7.3	8.1	185	254	1.0	1.7	0.60	1.20	70	210	280	840	-	-
34	1012	15.0	29.0	7.0	9.2	7.7	8.2	169	274	1.0	1.4	0.40	1.10	70	170	280	490	-	-
35	4164	18.0	33.0	6.9	9.0	7.6	8.2	189	1922	1.0	1.5	0.50	1.90	68	200	260	470	-	-
36	1867	10.0	10.0	9.3	9.3	7.6	7.6	3538	3538	1.0	1.0	0.32	0.32	2	2	1600	1600	2	2
37	2611	10.0	10.0	9.1	9.1	7.9	7.9	3637	3637	1.0	1.0	0.32	0.32	7	7	1600	1600	2	2
38	3872	6.0	13.0	8.8	9.3	7.5	8.4	1107	5072	1.0	1.4	0.32	1.42	2	1600	79	1600	2	4
39	4448	6.0	13.5	8.8	9.2	7.6	8.4	1166	5078	1.0	1.0	0.32	0.55	2	1600	46	1600	2	4
40	4449	6.5	13.0	8.9	9.3	7.7	8.4	1164	5087	1.0	1.0	0.32	1.41	2	240	70	1600	2	4
41	4452	7.0	13.5	8.7	9.3	7.6	8.3	1108	4635	1.0	1.0	0.32	1.22	2	1600	70	1600	2	2
42	3871	6.5	13.0	8.7	9.1	7.5	8.4	1111	5046	1.0	1.0	0.32	1.22	2	240	70	1600	2	2
43	3870	7.0	13.5	8.7	9.3	7.5	8.4	1159	4836	1.0	1.0	0.32	1.32	2	79	110	1600	2	2
44	1389	7.0	13.0	8.7	9.2	7.6	8.4	1124	4518	1.0	1.0	0.32	1.82	2	1600	130	1600	2	7
45	4453	7.0	13.5	8.7	9.2	7.4	8.3	1143	4559	1.0	1.0	0.32	1.22	2	110	130	1600	2	2
46	3869	8.0	14.9	8.5	9.1	7.2	8.3	257	2602	1.0	1.0	0.32	2.32	5	1600	350	1600	2	7
47	1086	8.0	14.5	8.6	9.1	7.5	8.5	1086	3840	1.0	1.0	0.32	1.32	8	1600	540	1600	2	94
48	1087	8.0	14.5	8.6	9.1	7.0	8.2	1020	2681	1.0	1.2	0.32	1.92	9	1600	1600	1600	2	220
49																			
50																			
51																			
52																			

Similarly do for all table.

Note

Do not copy abbreviation and note which given in different tables. Few examples are below

17	PRIMARY WATER QUALITY CRITERIA NOTIFIED UNDER E(P) RULES, 1986		> 5.0	6.5-8.5		< 3.0		< 2500 MPN/100ML		< 500 MPN/100ML

TABLE - 16.1: WATER QUALITY OF TRIBUTARY STREAMS - BUDHABALANGA AND SONO (2021)

STATION CODE	NAME OF MONITORING LOCATION	STATE NAME	TEMPERATURE (°C)		DISSOLVED OXYGEN (mg/L)		pH		CONDUCTIVITY (µmhos/cm)		BIO-CHEMICAL OXYGEN DEMAND		NITRATE (mg/ L)		FECAL COLIFORM (MPN/100 ML)		TOTAL COLIFORM (MPN/100 ML)		FECAL STREPTOCOCCI (MPN/100ML)	
			MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX
PRIMARY WATER QUALITY CRITERIA NOTIFIED UNDER E(P) RULES, 1986					> 5.0		6.5-8.5				< 3.0				< 2500 MPN/100ML				< 500 MPN/100ML	

NOTE:

ABOVE DATA IS AS PER THE INFORMATION PROVIDED BY THE SPCBS/PCCS THROUGH EWQDES PORTAL. FURTHER CLARIFICATIONS, IF ANY MAY PLEASE BE OBTAINED FROM RESPECTIVE SPCB/PCC. ALL THE INTER-STATE MONITORING LOCATIONS ARE MONITORED BY CPCB THROUGH HEAD OFFICE (DELHI) AND REGIONAL DIRECTORATES.
BDL VALUES FOR DISSOLVED OXYGEN (0.3 MG/L), CONDUCTIVITY (5 µMHOS/CM), BOD (1 MG/L), NITRATE-N + NITRITE-N (0.3 MG/L), FAECAL COLIFORM, TOTAL COLIFORM & FAECAL STREPTOCOCCI (1.8 MPN/100 ML).

ABBREVIATIONS

1. U/S - UP-STREAM
2. D/S - DOWNSTREAM
3. A/C - AFTER CONFLUENCE
4. B/C - BEFORE CONFLUENCE
5. BDL - BELOW DETECTION LIMIT
6. MG/L - MILLIGRAM PER LITER
7. MPN/ 100 ML - MOST PROBABLE NUMBER/ 100 MILLILITER
8. E (P) RULES1986 - ENVIRONMENT (PROTECTION) RULES, 1986

Save the file in excel and csv format.

Repeat the extraction of data proves from different software (small pdf)

Copy the converted data from both the file in one excel file in different sheet.

Subtract each corresponding numeric cell from sheet-1 to sheet -2 in sheet-3 using following Formula

='Table 1'!D4-'Table2'!D4

For string data ise following formula

=EXACT('Table 1'!D4,'Table 2'!C4)

Repeat the same process by another conversion software and subtract each corresponding cell and check for zero values

Sheet 1 sheet2 and sheet 3 is shown below.

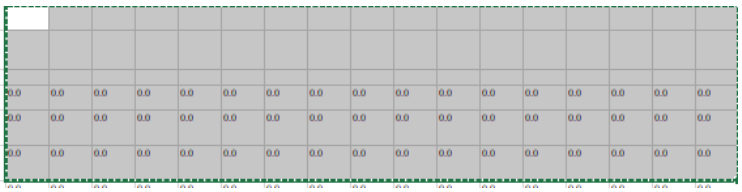
Table-1

Water Quality of Medium & Minor Rivers under NWMP-2021																			
STN Code	Name of Monitoring Location	State Name	Temperature °C		Dissolved Oxygen (mg/L)		pH		Conductivity (µmho/c		BOD (mg/L)		Nitrate N + Nitrite N(mg/L)		Fecal Coliform (MPN/100		Total Coliform (MPN/100ml	
			Min	Max	Min	Max	Min	Ma	Min	Max	Min	Ma	Min	Max	Min	Max	Min	Max
1448	RIVER NAGAVALI AT THOTAPALLI REGULATOR, TIZIANAGARAM	ANDHRA PRADESH	25.8	32.8	5.3	8.2	7.8	8.2	242	488	1.4	2.8	0.72	1.66	7	15	33	158
4351	RIVER NAGAVALI IMMEDIATE BORDER BETWEEN ANDHRA PRADESH & ORISSA STATES, NEAR SITALATAH, KHURDA [T],	ANDHRA PRADESH	24.8	32.8	5.2	8.3	6.7	8.5	248	642	1.2	2.5	0.52	3.82	11	21	158	248
4345	RIVER NAGAVALI NEAR NH- 16 BRIDGE, UPSTREAM [U/S] OF SRIKAKULAM TOWN, BEFORE CONFLUENCE OF TOWN	ANDHRA PRADESH	24.8	32.8	6.4	8.4	7.1	8.8	243	488	1.2	2.8	0.54	3.68	7	15	128	128
4347	RIVER NAGAVALI NEAR WATER PUMP HOUSE [KILLIPALEM] , DOWN STREAM [D/S] OF SRIKAKULAM TOWN, AFTER	ANDHRA PRADESH	24.8	32.8	6.8	7.5	7.1	8.4	338	648	1.4	2.3	0.55	4.82	11	15	158	248
4348	RIVER VAMSADHARA, IMMEDIATE BORDER	ANDHRA	22.8	28.8	5.8	8.3	6.3	8.5	224	448	1.5	2.2	0.42	2.46	3	21	128	248

Table-2

Water Quality of Medium & Minor Rivers under NWMP-2021																			
STN Code	Name of Monitoring Location	State Name	Temperature °C		Dissolved Oxygen (mg/L)		pH		Conductivity (µmho/c		BOD (mg/L)		Nitrate N + Nitrite N(mg/L)		Fecal Coliform (MPN/100		Total Coliform (MPN/100ml	
			Min	Max	Min	Max	Min	Ma	Min	Max	Min	Ma	Min	Max	Min	Max	Min	Max
1448	RIVER NAGAVALI AT THOTAPALLI REGULATOR, TIZIANAGARAM	ANDHRA PRADESH	25.8	32.8	5.3	8.2	7.8	8.2	242	488	1.4	2.8	0.72	1.66	7	15	33	158
4351	RIVER NAGAVALI IMMEDIATE BORDER BETWEEN ANDHRA PRADESH & ORISSA STATES, NEAR SITALATAH, KHURDA [T],	ANDHRA PRADESH	24.8	32.8	5.2	8.3	6.7	8.5	248	642	1.2	2.5	0.52	3.82	11	21	158	248
4345	RIVER NAGAVALI NEAR NH- 16 BRIDGE, UPSTREAM [U/S] OF SRIKAKULAM TOWN, BEFORE CONFLUENCE OF TOWN	ANDHRA PRADESH	24.8	32.8	6.4	8.4	7.1	8.8	243	488	1.2	2.8	0.54	3.68	7	15	128	128

Table-3



If all the cell has zero . it means the conversion is proper. If there is error in any cell then check both the cell. It may be there some special character like "-" "n.a.". Check the values from Original pdf and make correction.

Data Cleaning

- Duplicate removal
- Special character
- Noise removal
- Fixing Data type
- Standardize capitalization

Now the converted data is available. Sometime in the conversion process additional row may occurs. This type error is detected in the validation process. It may possible, it is not detected in validation process if that is not detected. Sometime in the appending or adding the data in the table it may possible some duplicate row may be a added.

It is also possible in original file there may be some special character for foot note , Some time blank space is also represented by in the "*", "N.A","$$","!"

Duplicate Removal

- First find out which column has unique no.. In present case station id is unique for each row . here we compare by two consecutive row by following formula
- **EXACT(A4, A5)** if the result is Flse , it means these two rows are not dulicate .

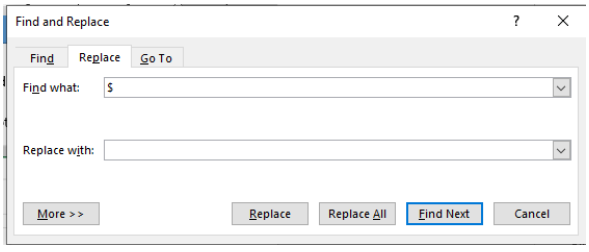
	A	B	C
1		Water Quality of Me	
2	STN Code	Name of Monitoring Location	State Name
3			
4	1448	RIVER NAGAVALI AT THOTAPALLI REGULATOR, VIZIANAGARAM	ANDHRA PRADESH
5	4351	RIVER NAGAVALI IMMEDIATE BORDER BETWEEN ANDHRA PRADESH & ORISSA STATES, NEAR SIVALAYAM, KUNERU (V),	ANDHRA PRADESH
6	4346	RIVER NAGAVALI NEAR NH- 16 BRIDGE, UPSTREAM (U/S) OF SRIKAKULAM TOWN, BEFORE CONFLUENCE OF TOWN SEWAGE	ANDHRA PRADESH
7	4347	RIVER NAGAVALI NEAR WATER PUMP HOUSE (KILLIPALEM) , DOWN STREAM (D/S) OF SRIKAKULAM TOWN, AFTER	ANDHRA PRADESH
8	4348	RIVER VAMSADHARA, IMMEDIATE BORDER BETWEEN ANDHRA PRADESH & ORISSA STATES, NEAR LALITHAMBA	ANDHRA PRADESH

EXACT(A4, A5)

V
FALSE
FALSE
FALSE
FALSE
FALSE

Special charterer Removal

Use replace statement to remove noise and special characters



Fixing data type

Before finalizing the data it is needed to check the data type of data.

Weather data is Integer, floating, Number of decimal place, Length of string.

In CPCB data base temperature is numeric with single decimal. Station Id as String of 10 character. The precision of data should be decided on the original data which has been transformed.

Data Validation

- Checking of extreme data
- Checking of Missing data
- Checking, fixing and removing outlier
- Mapping the data (Spatial data)

Checking of Extreme data

In excel, we can find out the extreme values of any row or any column by following formula

Subtotal (4, starting cell of selected Column: last cell of Selected Column)

It will give the maximum value selected column

Similarly

Subtotal (5, starting cell of selected Column: last cell of Selected Column) will give the minimum value

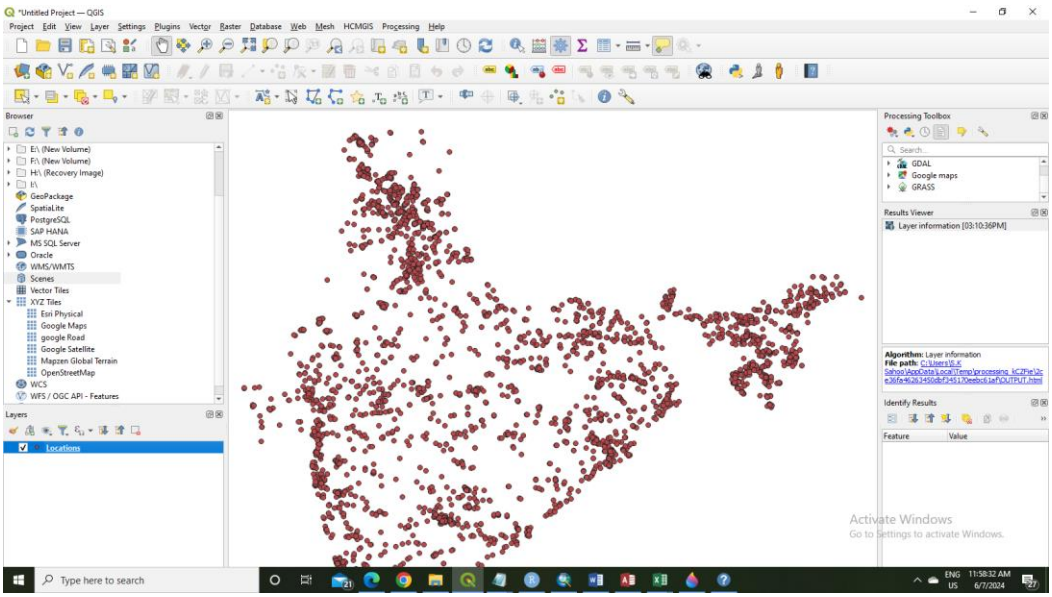
By finding out the maximum value and minimum value of column, one can make judgment based on the specific characteristic parameter. i.e the temperature of water bodies generally varies between 4°C to 40°C. of the max value and Min value beyond this range, the re check it from the original data/ PDF file.

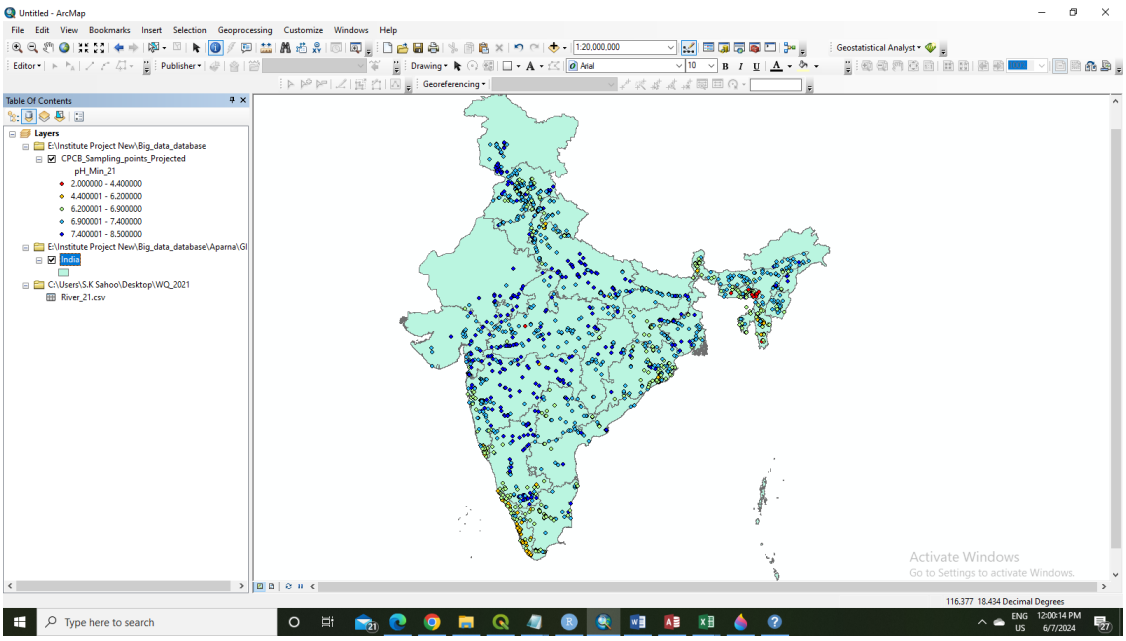
We can also check that Max temperature should be more then minimum temperature . by subtractin t_max by t_min. similar analysis should also perform on other parameters

If the some values is unexplainable then it may remove.

Mapping the data

Sometime the data is spatial in nature in that case the values of X and Y or Latitude and longitude then compare the values weather it is coming in the targeted location. Location data (X, Y or latitude/ longitude) should be projected and check whether the data point or shapes is projecting within the specific boundary. Here we can use open source GIS software (QGIS) or commercial software (ARCGIS)





Data storing format

There are so many formats for the Data storage. We have to select the data format such a manner that is acceptable by many software and simple. For CPCB data, we are saving the data in Coma separated Value (CSV) format