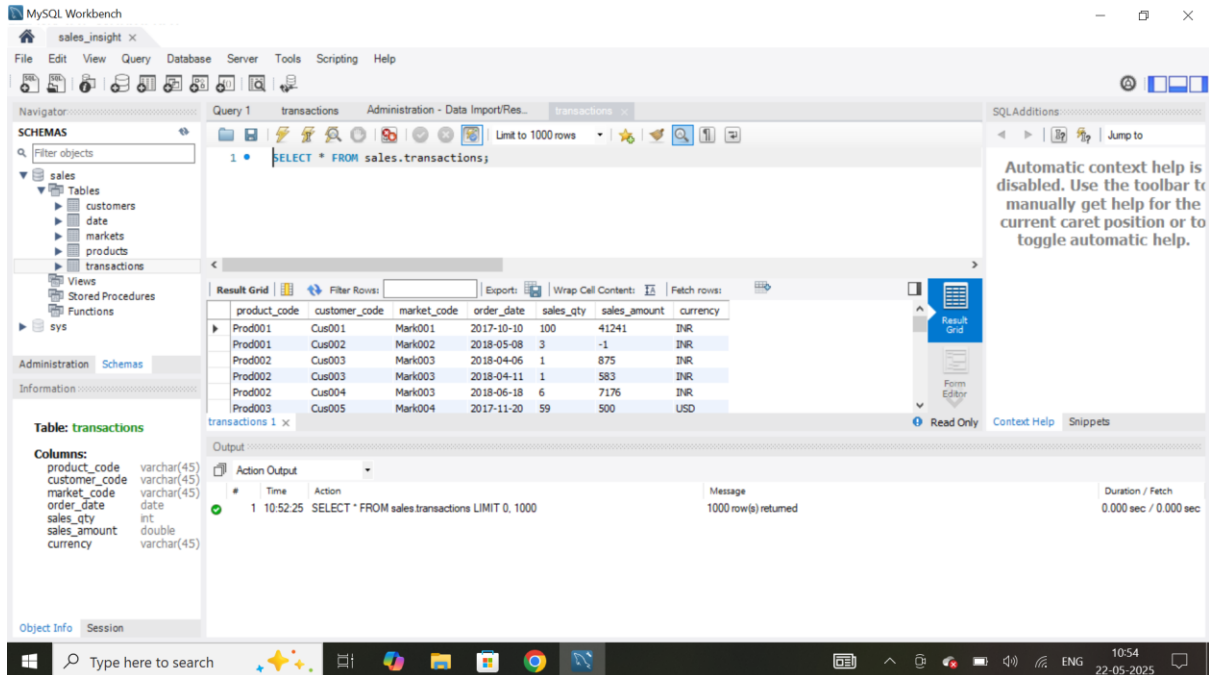


Data Preprocessing & Exploratory Data Analysis (EDA)

We have a company's sales insight and we are doing analysis on that data

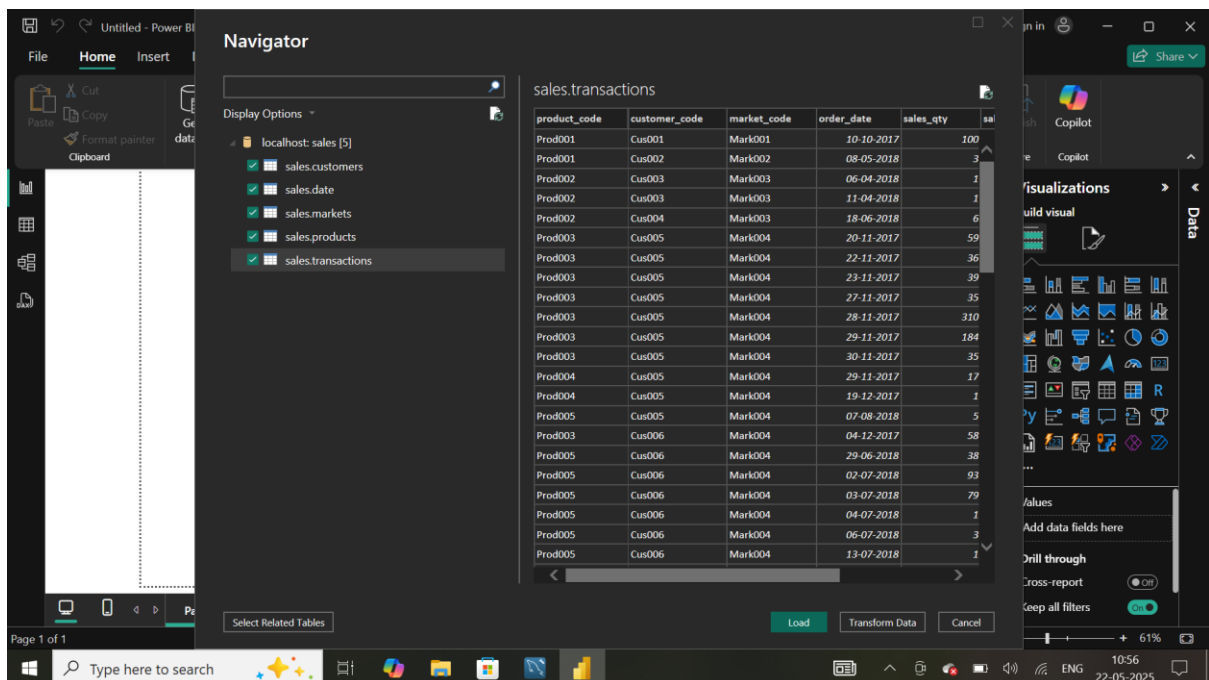


The screenshot shows the MySQL Workbench interface. The 'Query 1' tab is active, displaying a SQL query: `SELECT * FROM sales.transactions;`. The 'Result Grid' shows the first 10 rows of the 'sales.transactions' table. The columns are: product_code, customer_code, market_code, order_date, sales_qty, sales_amount, and currency. The data is as follows:

product_code	customer_code	market_code	order_date	sales_qty	sales_amount	currency
Prod001	Cus001	Mark001	2017-10-10	100	41241	INR
Prod001	Cus002	Mark002	2018-05-08	3	-1	INR
Prod002	Cus003	Mark003	2018-04-06	1	875	INR
Prod002	Cus003	Mark003	2018-04-11	1	583	INR
Prod002	Cus004	Mark004	2018-06-18	6	7176	INR
Prod003	Cus005	Mark004	2017-11-20	59	500	USD

The 'Output' tab shows the execution details: 10:52:25, Action: SELECT * FROM sales.transactions LIMIT 0, 1000, Message: 1000 row(s) returned, Duration / Fetch: 0.000 sec / 0.000 sec.

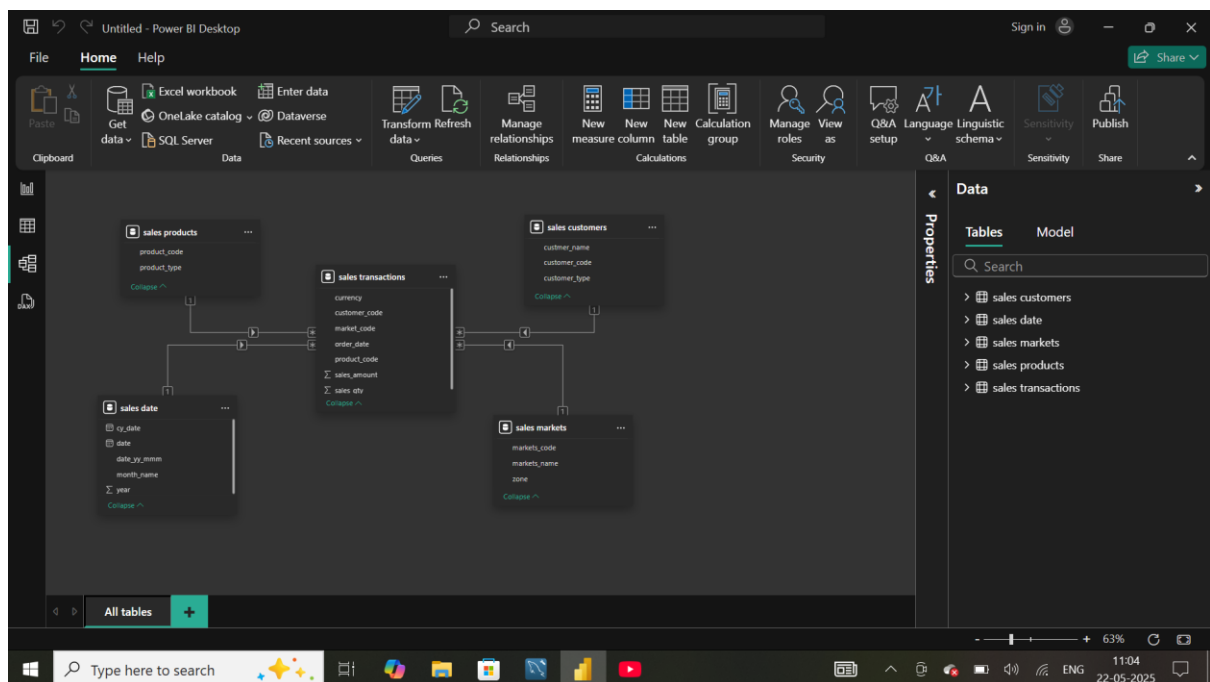
Now we will upload the data in Power BI



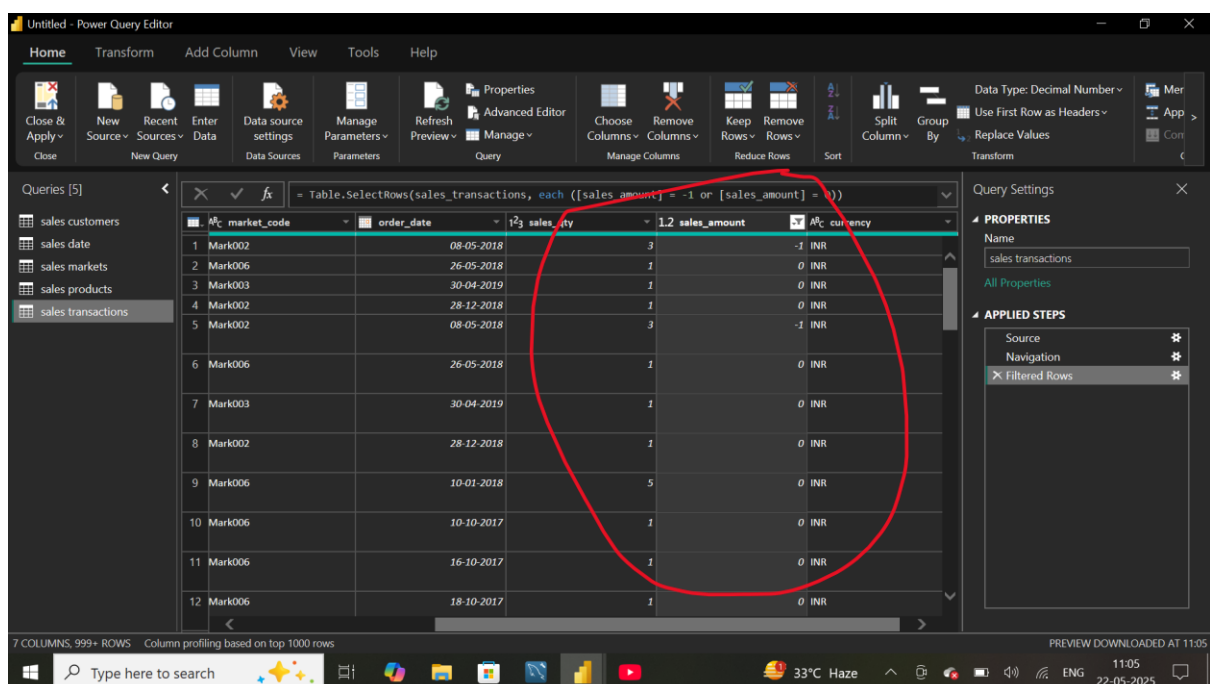
The screenshot shows the Microsoft Power BI Desktop interface. The 'Navigator' pane on the left lists the tables in the 'localhost: sales [5]' database. The 'sales.transactions' table is selected. The main view displays the data from the 'sales.transactions' table, showing columns: product_code, customer_code, market_code, order_date, sales_qty, and sales_amount. The data is as follows:

product_code	customer_code	market_code	order_date	sales_qty	sales_amount
Prod001	Cus001	Mark001	10-10-2017	100	41241
Prod001	Cus002	Mark002	08-05-2018	3	-1
Prod002	Cus003	Mark003	06-04-2018	1	875
Prod002	Cus003	Mark003	11-04-2018	1	583
Prod002	Cus004	Mark003	18-06-2018	6	7176
Prod003	Cus005	Mark004	20-11-2017	59	500
Prod003	Cus005	Mark004	22-11-2017	36	
Prod003	Cus005	Mark004	23-11-2017	39	
Prod003	Cus005	Mark004	27-11-2017	35	
Prod003	Cus005	Mark004	28-11-2017	310	
Prod003	Cus005	Mark004	29-11-2017	184	
Prod003	Cus005	Mark004	30-11-2017	35	
Prod004	Cus005	Mark004	29-11-2017	17	
Prod004	Cus005	Mark004	19-12-2017	1	
Prod005	Cus005	Mark004	07-08-2018	5	
Prod005	Cus006	Mark004	04-12-2017	58	
Prod005	Cus006	Mark004	29-06-2018	38	
Prod005	Cus006	Mark004	02-07-2018	93	
Prod005	Cus006	Mark004	03-07-2018	79	
Prod005	Cus006	Mark004	04-07-2018	1	
Prod005	Cus006	Mark004	06-07-2018	3	
Prod005	Cus006	Mark004	13-07-2018	1	

Since we have uploaded the data, now we will create relation between the tables

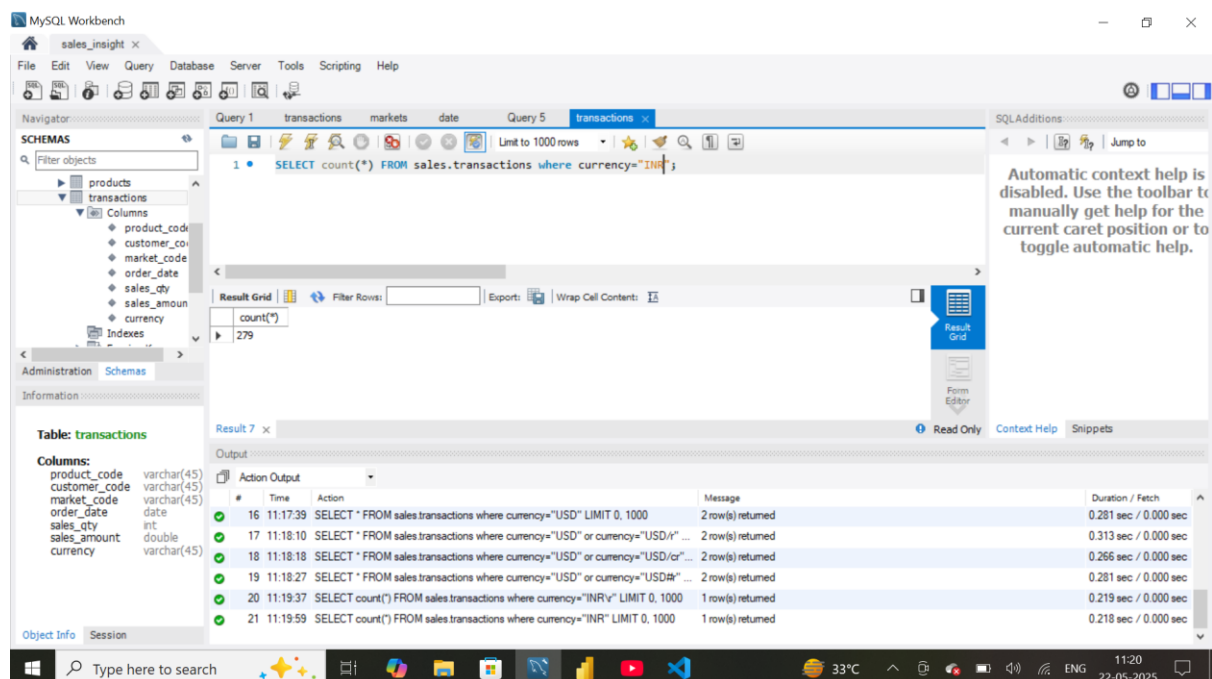
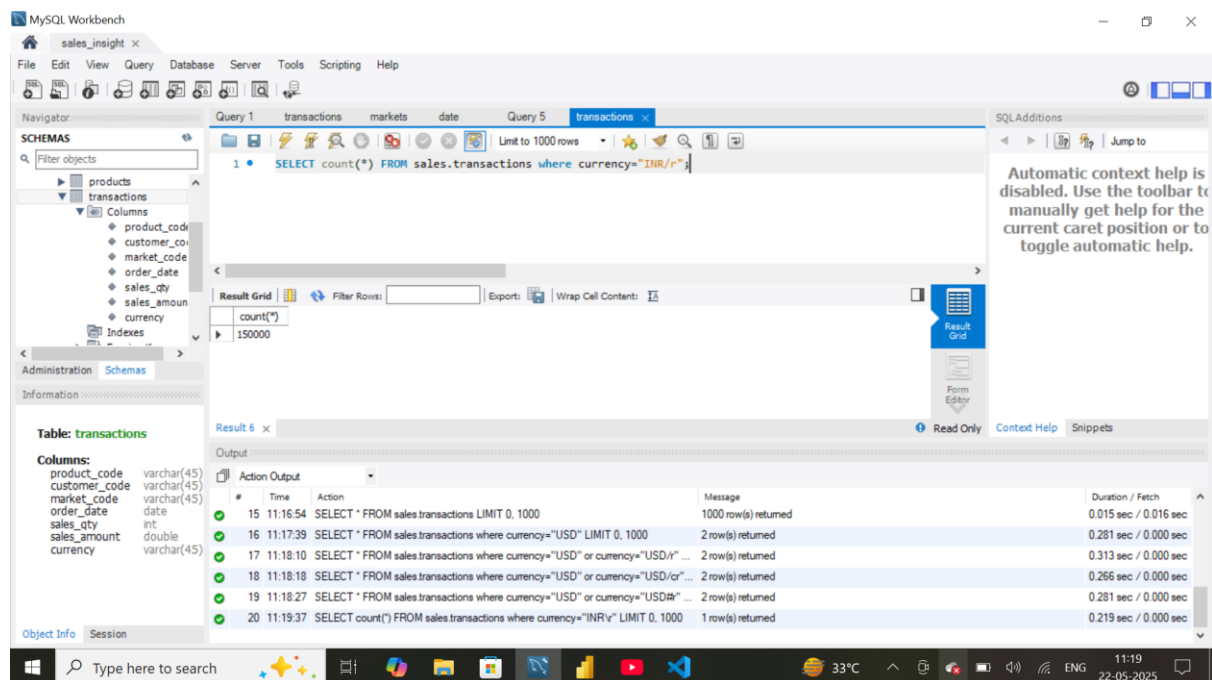


Now are tables are having relational schema and now we will do data cleaning



As you can see 0 and -1 in the sales_amount column, so we will remove these values using **"= Table.SelectRows(sales_transactions, each ([sales_amount] <> -1 and [sales_amount] <> 0))"** This filters the column and all the values are removed

Now we noticed that in currency column some currency are duplicate



As you can see in above photos that INR/r has more values as compare to INR so will remove INR using **“= Table.SelectRows("#Remove sales_amount= 0/-1", each ([currency] = "INR" or [currency] = "USD"))”** This will remove the rows have INR and USD currency which are duplicate.

Now we will convert USD currency to normal amount using “= **Table.AddColumn(#"Removing duplicity", "norm_amount", each if Text.Trim([currency]) = "USD" then [sales_amount]*75 else [sales_amount])**”

Query: = Table.AddColumn(#"Removing duplicity", "norm_amount", each if Text.Trim([currency]) = "USD" then [sales_amount]*75 else [sales_amount])

order_date	sales_qty	sales_amount	currency	norm_amount
10-10-2017	100	41241	INR	412
06-04-2018	1	875	INR	8
11-04-2018	1	583	INR	5
18-06-2018	6	7176	INR	71
20-11-2017	59	500	USD	375
22-11-2017	36	250	USD	187
23-11-2017	39	21412	INR	214
27-11-2017	35	19213	INR	192
28-11-2017	310	170185	INR	1701
29-11-2017	184	101194	INR	1011
30-11-2017	35	19213	INR	192
29-11-2017	17	9426	INR	94
19-12-2017	1	218	INR	2
07-08-2018	5	3093	INR	30
04-12-2017	58	30306	INR	303
29-06-2018	38	52319	INR	523
02-07-2018	93	126296	INR	1262
03-07-2018	79	107500	INR	1075
04-07-2018	1	273	INR	2

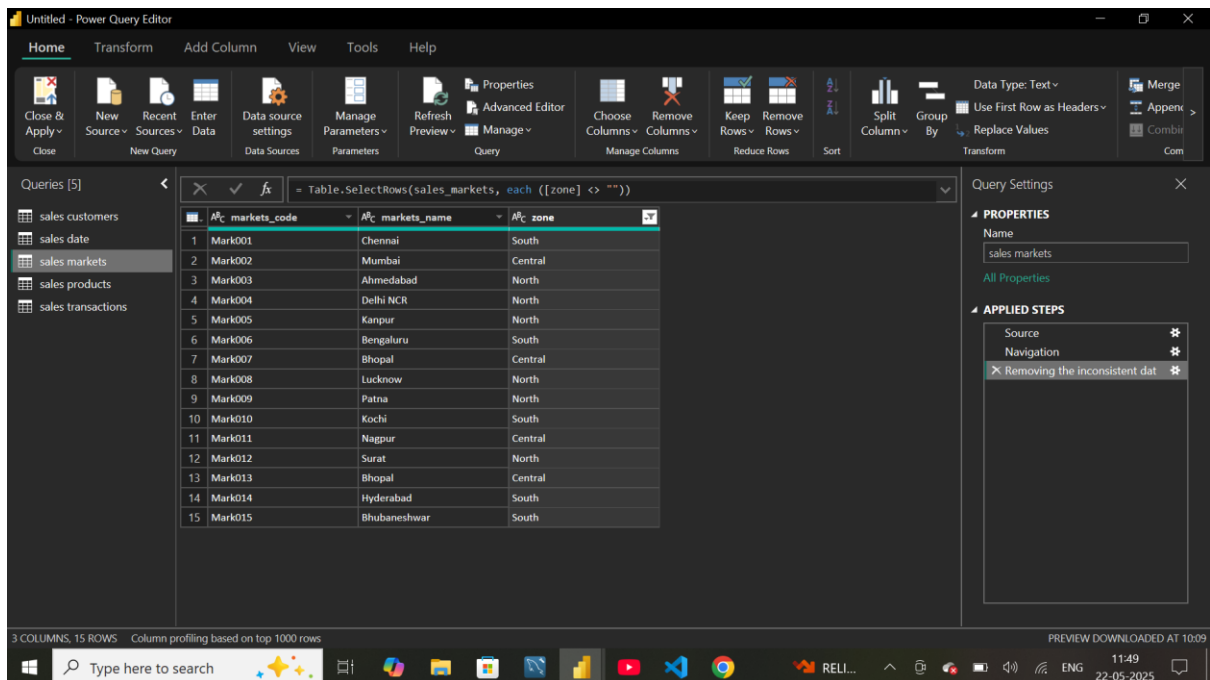
So finally our transactions table is ready for analysis and our other tables are also correct, but in sales.markets we some inconsistent values

Query: = Source[[Schema="sales",Item="markets"]][data]

markets_code	markets_name	zone
Mark001	Chennai	South
Mark002	Mumbai	Central
Mark003	Ahmedabad	North
Mark004	Delhi NCR	North
Mark005	Kanpur	North
Mark006	Bengaluru	South
Mark007	Bhopal	Central
Mark008	Lucknow	North
Mark009	Patna	North
Mark010	Kochi	South
Mark011	Nagpur	Central
Mark012	Surat	North
Mark013	Bhopal	Central
Mark014	Hyderabad	South
Mark015	Bhubaneswar	South
Mark097	New York	
Mark999	Paris	

We will remove these rows as they are of no use and cause error in analysing

“= Table.SelectRows(sales_markets, each ([zone] <> ""))” This will remove the rows



The screenshot shows the Power Query Editor interface. The formula bar at the top displays the M code: `= Table.SelectRows(sales_markets, each ([zone] <> ""))`. Below the formula bar, a table with 15 rows is visible. The columns are `markets_code`, `markets_name`, and `zone`. The data is as follows:

	markets_code	markets_name	zone
1	Mark001	Chennai	South
2	Mark002	Mumbai	Central
3	Mark003	Ahmedabad	North
4	Mark004	Delhi NCR	North
5	Mark005	Kanpur	North
6	Mark006	Bengaluru	South
7	Mark007	Bhopal	Central
8	Mark008	Lucknow	North
9	Mark009	Patna	North
10	Mark010	Kochi	South
11	Mark011	Nagpur	Central
12	Mark012	Surat	North
13	Mark013	Bhopal	Central
14	Mark014	Hyderabad	South
15	Mark015	Bhubaneshwar	South

The right-hand pane shows the 'Query Settings' for 'sales markets'. The 'APPLIED STEPS' list includes 'Source', 'Navigation', and 'Removing the inconsistent data'.

So now are whole data is correct and cleaned for analysis.