

## SQL Document:

Creating the table and importing content into the table

**/\* Creating Table \*/**

**Create table MIO(**

**Typeofsales varchar,**

**Patient\_ID Bigint,**

**Specialisation varchar,**

**Dept varchar,**

**Dateofbill varchar,**

**Quantity int,**

**ReturnQuantity int,**

**Final\_Cost DOUBLE PRECISION,**

**Final\_Sales DOUBLE PRECISION,**

**RtnMRP DOUBLE PRECISION,**

**Formulation varchar,**

**DrugName varchar,**

**SubCat varchar,**

**SubCat1 varchar**

**)**

To import data into the created MIO table. Right Click on table Name on left side of panel -> select import and export Data -> choose the csv file to import -> cross check the column names in column tab -> Click ok

Selecting or Viewing data from the table:

**select \* from MIO**

	typesales character varying	patient_id bigint	specialisation character varying	dept character varying	dateofbill character varying	quantity integer	returnquantity integer	final_cost double precision	final_sales double precision	rtmr doul
1	Sale	12018098765	Specialisation6	Department1	06-01-2022	1	0	55.406	59.26	
2	Sale	12018103897	Specialisation7	Department1	7/23/2022	1	0	768.638	950.8	
3	Sale	12018101123	Specialisation2	Department3	6/23/2022	1	0	774.266	4004.214	
4	Sale	12018079281	Specialisation40	Department1	3/17/2022	2	0	40.798	81.044	
5	Sale	12018117928	Specialisation5	Department1	12/21/2022	1	0	40.434	40.504	
6	Return	12018103662	Specialisation2	Department1	7/15/2022	0	8	47.902	0	
7	Sale	12018097585	Specialisation2	Department1	5/22/2022	1	0	41.862	42.218	
8	Sale	12018077721	Specialisation4	Department1	01-12-2022	3	0	60.026	142.752	
9	Sale	12018096500	Specialisation4	Department2	8/24/2022	2	0	49.856	94	
10	Sale	12018071649	Specialisation4	Department1	8/31/2022	1	0	258.86	319.8	
11	Sale	12018074904	Specialisation7	Department1	10-04-2022	2	0	114.502	200.4	
Total rows: 1000 of 14218    Query complete 00:00:00.125    Ln 21, Col 1										

Total Data in dataset before cleaning 14218. The data consists of null values in some fields.

Creating view: Creating view to do feature engineering. We are using this view for further Analysis

--Create view from MIO table

create view MioView as

select \* from MIO

	typesales character varying	patient_id bigint	specialisation character varying	dept character varying	dateofbill character varying	quantity integer	returnquantity integer	final_cost double precision	final_sales double precision	rtmr doul
1	Sale	12018098765	Specialisation6	Department1	06-01-2022	1	0	55.406	59.26	
2	Sale	12018103897	Specialisation7	Department1	7/23/2022	1	0	768.638	950.8	
3	Sale	12018101123	Specialisation2	Department3	6/23/2022	1	0	774.266	4004.214	
4	Sale	12018079281	Specialisation40	Department1	3/17/2022	2	0	40.798	81.044	
5	Sale	12018117928	Specialisation5	Department1	12/21/2022	1	0	40.434	40.504	
6	Return	12018103662	Specialisation2	Department1	7/15/2022	0	8	47.902	0	
7	Sale	12018097585	Specialisation2	Department1	5/22/2022	1	0	41.862	42.218	
8	Sale	12018077721	Specialisation4	Department1	01-12-2022	3	0	60.026	142.752	
9	Sale	12018096500	Specialisation4	Department2	8/24/2022	2	0	49.856	94	
10	Sale	12018071649	Specialisation4	Department1	8/31/2022	1	0	258.86	319.8	
11	Sale	12018074904	Specialisation7	Department1	10-04-2022	2	0	114.502	200.4	
Total rows: 1000 of 14218    Query complete 00:00:00.125    Ln 21, Col 1										

Data Preprocessing:

Describing Dataset:

/\*Describing Dataset \*/

SELECT

column\_name,

data\_type,

character\_maximum\_length,

is\_nullable

FROM

information\_schema.columns

WHERE

table\_name = 'mio' AND table\_schema = 'public'

	column_name name	data_type character varying	character_maximum_length integer	is_nullable character varying (3)
1	typesales	character varying	[null]	YES
2	patient_id	bigint	[null]	YES
3	specialisation	character varying	[null]	YES
4	dept	character varying	[null]	YES
5	dateofbill	character varying	[null]	YES
6	quantity	integer	[null]	YES
7	returnquantity	integer	[null]	YES
8	final_cost	double precision	[null]	YES
9	final_sales	double precision	[null]	YES
10	rtnmrp	double precision	[null]	YES
11	formulation	character varying	[null]	YES
Total rows: 14 of 14		Query complete 00:00:00.069		

### Checking duplicate values in patient id fields:

select Dateofbill,patient\_id, final\_sales from MioVieworder by patient\_id;

	dateofbill character varying	patient_id bigint	final_sales double precision
1	08-06-2022	12017998218	447.2
2	02-05-2022	12017998218	81
3	7/15/2022	12017998261	160.51
4	11-08-2022	12017998278	43.33
5	11-09-2022	12017998278	415.89
6	12/16/2022	12017998291	0
7	5/29/2022	12017998321	63.9
8	7/28/2022	12017998321	138.63
9	5/15/2022	12017998321	331
10	5/27/2022	12017998321	173.332
11	5/22/2022	12017998321	0
Total rows: 1000 of 14218		Query complete 00:00:00.085	

Checking Null values: checking if the column contains null values

-- checking null values in a column

select \* from MioView where patient\_id is NULL

select \* from MioView where specialisation is NULL

select \* from MioView where dept is NULL

select \* from MioView where dateofbill is NULL

select \* from MioView where quantity is NULL

select \* from MioView where returnquantity is NULL

select \* from MioView where final\_cost is NULL

select \* from MioView where final\_sales is NULL

select \* from MioView where rtnmrp is NULL

select \* from MioView where formulation is NULL -- contains null value

	Lcost ble precision	final_sales double precision	rtnmrp double precision	formulation character varying	drugname character varying	subcat character varying	subcat1 character varying
1	64.864	0	96.8	[null]	MULTIPLE ELECTROLYTES 500ML IVF	IV FLUIDS, ELECTROLYTES, TPN	INTRAVENOUS & i
2	52.544	181.12	0	[null]	POTASSIUM CHLORIDE 150MG	INJECTIONS	INTRAVENOUS & i
3	57.408	0	61.76	[null]	CALCIUM 250MG + VITAMIN D3 125IU	TABLETS & CAPSULES	VITAMINS & MINE
4	46.212	0	46.986	[null]	DEXTROSE 10%W/V 500ML IVF	IV FLUIDS, ELECTROLYTES, TPN	INTRAVENOUS & i
5	114.592	0	290.4	[null]	MULTIPLE ELECTROLYTES 500ML IVF	IV FLUIDS, ELECTROLYTES, TPN	INTRAVENOUS & i
6	58.704	0	121.6	[null]	[null]	[null]	[null]
7	70.912	0	151.8	[null]	DOXYCYCLINE 100MG INJ	INJECTIONS	ANTH-INFECTIVES
8	51.122	0	55.86	[null]	SODIUM CHLORIDE 0.9%	IV FLUIDS, ELECTROLYTES, TPN	INTRAVENOUS & i
9	46.272	0	90.56	[null]	POTASSIUM CHLORIDE 150MG	INJECTIONS	INTRAVENOUS & i
10	62.266	0	782	[null]	[null]	[null]	[null]

select \* from MioView where drugname is NULL -- contains null values

	quantity integer	returnquantity integer	final_cost double precision	final_sales double precision	rtnmrp double precision	formulation character varying	drugname character varying	subcat character varying	subcat1 character varying
1	1	0	49.352	60.8	0	Form1	[null]	[null]	[null]
2	2	0	40.34	81.1	0	Form1	[null]	[null]	[null]
3	2	0	40.34	81.1	0	Form1	[null]	[null]	[null]
4	1	0	49.956	62.8	0	Form1	[null]	[null]	[null]
5	4	0	77.408	243.2	0	Form1	[null]	[null]	[null]
6	1	0	49.352	60.8	0	Form1	[null]	[null]	[null]
7	4	0	79.828	251.2	0	Form1	[null]	[null]	[null]
8	10	0	41.702	405	0	Form1	[null]	[null]	[null]
9	1	0	49.956	62.8	0	Form1	[null]	[null]	[null]
10	0	2	58.704	0	121.6	[null]	[null]	[null]	[null]

Total rows: 1000 of 1668    Query complete 00:00:00.111    Ln 48, Col 1

select \* from MioView where subcat is NULL -- contains null values

Data Output Messages Notifications										
	quantity integer	returnquantity integer	final_cost double precision	final_sales double precision	rtmnp double precision	formulation character varying	drugname character varying	subcat character varying	subcat1 character varying	
1	1	0	49.352	60.8	0	Form1	[null]	[null]	[null]	
2	2	0	40.34	81.1	0	Form1	[null]	[null]	[null]	
3	2	0	40.34	81.1	0	Form1	[null]	[null]	[null]	
4	1	0	49.956	62.8	0	Form1	[null]	[null]	[null]	
5	4	0	77.408	243.2	0	Form1	[null]	[null]	[null]	
6	1	0	49.352	60.8	0	Form1	[null]	[null]	[null]	
7	4	0	79.828	251.2	0	Form1	[null]	[null]	[null]	
8	10	0	41.702	405	0	Form1	[null]	[null]	[null]	
9	1	0	49.956	62.8	0	Form1	[null]	[null]	[null]	
10	0	2	58.704	0	121.6	[null]	[null]	[null]	[null]	
Total rows: 1000 of 1668 Query complete 00:00:00.111 Ln 49, Co										

select \* from MioView where subcat1 is NULL -- contains null values

	formulation character varying	drugname character varying	subcat character varying	subcat1 character varying	
1	Form1	[null]	[null]	[null]	
2	Form1	[null]	[null]	[null]	
3	Form1	[null]	[null]	[null]	
4	Form1	[null]	[null]	[null]	
5	Form1	[null]	[null]	[null]	
6	Form1	[null]	[null]	[null]	
7	Form1	[null]	[null]	[null]	
8	Form1	[null]	[null]	[null]	
9	Form1	[null]	[null]	[null]	
10	[null]	[null]	[null]	[null]	
Total rows: 1000 of 1692 Query complete 00:00:00.179 Ln 50, Col 1					

From the above queries we can see Columns formulation, Drugname,subcat,subcat1 contains null value so we need to replace the null value with NA Value so it won't affect the remaining data while performing EDA

### Replacing Null Values with NA

-- replace NULL values in Formulation column

Update MioView set formulation = COALESCE(formulation, 'NA');

-- replace NULL values in drugname,subcat,subcat1 column

Update MioView

set drugname = COALESCE(drugname, 'NA'),

subcat = COALESCE(subcat,'NA'),

subcat1 = COALESCE(subcat1,'NA');

--selecting Patient\_id columns where drugname is NA

select patient\_id, drugname from MioView where drugname = 'NA'

Data Output		Messages	Notifications
	patient_id bigint	drugname character varying	
1	12018086686	NA	
2	12018111593	NA	
3	12018105512	NA	
4	12018095122	NA	
5	12018073513	NA	
6	12018106818	NA	
7	12018096947	NA	
8	12018120691	NA	
9	12018090142	NA	
10	12018079574	NA	
11	12018104322	NA	
Total rows: 1000 of 1668		Query complete	

Counting NA Values from columns:

select

count(CASE WHEN formulation = 'NA' then 1 END) as Formulation\_NA\_Count,

count(CASE WHEN drugname = 'NA' then 1 END) as Drugname\_NA\_Count,

count(CASE WHEN subcat = 'NA' then 1 END) as Subcat\_NA\_count,

count(case when subcat1 = 'NA' then 1 END) as Subcat1\_NA\_count

from MioView

	formulation_na_count bigint	drugname_na_count bigint	subcat_na_count bigint	subcat1_na_count bigint
1	653	1668	1668	1692

From the above table formulation contains 653 null values and drugname and subcat both contains 1668 null values each and subcat1 contains 1692

## Data Formatting:

From the observation in date column, the format is inconsistent. 08-06-2022 and 7/15/2022 are the two formats we can see in the date column. So we are changing the format uniformly to 7/15/2022

-- Formatting date format

Update MioView

```
Set dateofbill = TO_CHAR(TO_DATE(dateofbill, 'MM/DD/YY'), 'DD/MM/YY')
```

```
WHERE dateofbill IS NOT NULL AND dateofbill != '';
```

```
select * from MioView;
```

	typesales character varying	patient_id bigint	specialisation character varying	dept character varying	dateofbill character varying	quantity integer	returnquantity integer	final_cost double precision	final_sales double precision	rtmr doul
1	Sale	12018111758	Specialisation5	Department1	15/10/22	10	0	41.702	405.5	
2	Sale	12018084992	Specialisation2	Department3	28/01/22	1	0	838	2444	
3	Sale	12018078536	Specialisation4	Department1	17/02/22	1	0	51.442	53.828	
4	Sale	12018101963	Specialisation11	Department2	02/11/22	3	0	47.66	129.468	
5	Sale	12018089694	Specialisation7	Department1	13/03/22	3	0	97.12	415.89	
6	Sale	12018109715	Specialisation33	Department1	02/09/22	4	0	43.942	171.624	
7	Return	12018108216	Specialisation3	Department1	17/08/22	0	1	51.85	0	
8	Sale	12018079466	Specialisation7	Department1	28/02/22	50	0	105.49	2300	
9	Sale	12018051282	Specialisation3	Department1	30/04/22	1	0	42.464	49.984	
10	Sale	12018110286	Specialisation14	Department1	18/09/22	1	0	49.352	60.8	
Total rows: 1000 of 14218    Query complete 00:00:00.129    Ln 75, Col 1										

## Feature engineering:

Create column with month name from date and add that column and remaining cleaned data from MioView to other view name MioView1



-- adding a new column in other view

create view MioView1 as

```
select *,To_CHAR(to_date(dateofbill,'dd/mm/yy'),'Mon') as month from MioView
```

--selecting dateofbill and month values from MioView1

```
select dateofbill,month from MioView1
```

	<b>dateofbill</b> character varying 	<b>month</b> text 
1	15/10/22	Oct
2	28/01/22	Jan
3	17/02/22	Feb
4	02/11/22	Nov
5	13/03/22	Mar
6	02/09/22	Sep
7	17/08/22	Aug
8	28/02/22	Feb
9	30/04/22	Apr
10	18/09/22	Sep
11	01/08/22	Aug
Total rows: 1000 of 14218		Query complete 00:00:00.086