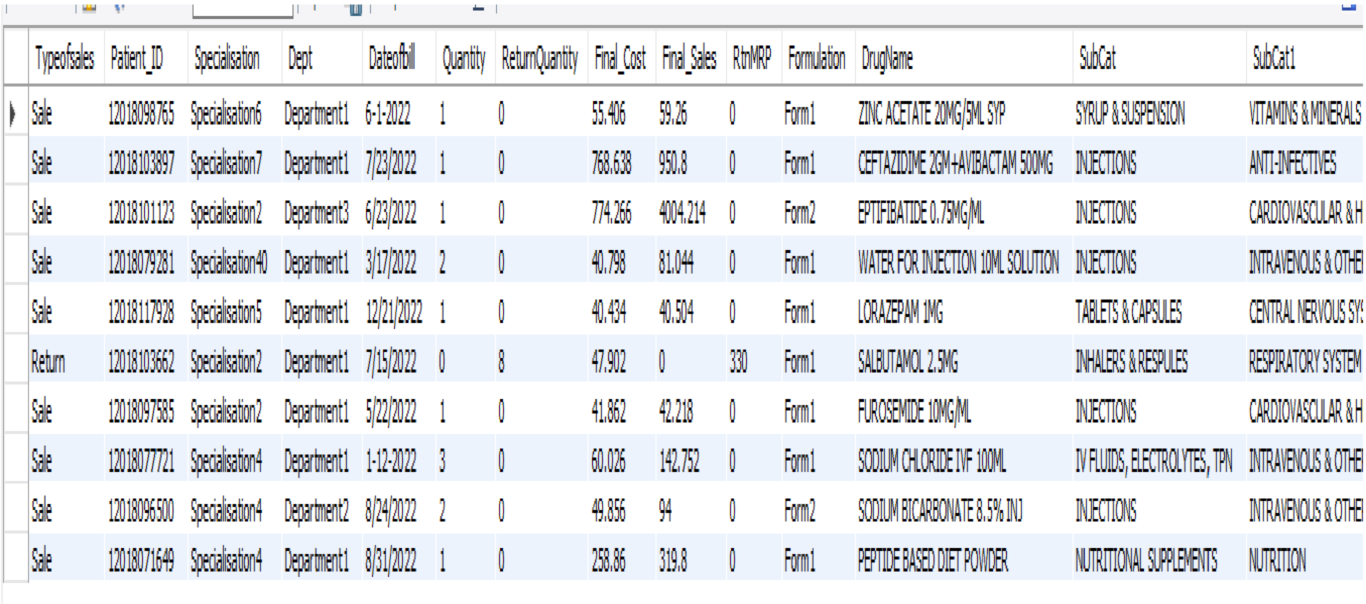
SQL Preprocessing Querries

Query:

select * from projectfinaldata limit 10;

Output:

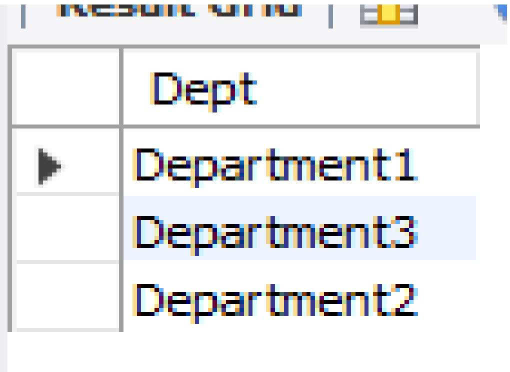


Typeofsales	Patient_ID	Specialisation	Dept	Dateofbill	Quantity	ReturnQuantity	Final_Cost	Final_Sales	RtnMRP	Formulation	DrugName	SubCat	SubCat1
Sale	12018098765	Specialisation6	Department1	6-1-2022	1	0	55.406	59.26	0	Form1	ZINC ACETATE 20MG/5ML SYP	SYRUP & SUSPENSION	VITAMINS & MINERALS
Sale	12018103897	Specialisation7	Department1	7/23/2022	1	0	768.638	950.8	0	Form1	CEFTAZIDIME 2GM+AVIBACTAM 500MG	INJECTIONS	ANTI-INFECTIVES
Sale	12018101123	Specialisation2	Department3	6/23/2022	1	0	774.266	4004.214	0	Form2	EPTIFIBATIDE 0.75MG/ML	INJECTIONS	CARDIOVASCULAR & H
Sale	12018079281	Specialisation40	Department1	3/17/2022	2	0	40.798	81.044	0	Form1	WATER FOR INJECTION 10ML SOLUTION	INJECTIONS	INTRAVENOUS & OTHE
Sale	12018117928	Specialisation5	Department1	12/21/2022	1	0	40.434	40.504	0	Form1	LORAZEPAM 1MG	TABLETS & CAPSULES	CENTRAL NERVOUS SY
Return	12018103662	Specialisation2	Department1	7/15/2022	0	8	47.902	0	330	Form1	SALBUTAMOL 2.5MG	INHALERS & RESPULES	RESPIRATORY SYSTEM
Sale	12018097585	Specialisation2	Department1	5/22/2022	1	0	41.862	42.218	0	Form1	FUROSEMIDE 10MG/ML	INJECTIONS	CARDIOVASCULAR & H
Sale	12018077721	Specialisation4	Department1	1-12-2022	3	0	60.026	142.752	0	Form1	SODIUM CHLORIDE IVF 100ML	IV FLUIDS, ELECTROLYTES, TPN	INTRAVENOUS & OTHE
Sale	12018096500	Specialisation4	Department2	8/24/2022	2	0	49.856	94	0	Form2	SODIUM BICARBONATE 8.5% INJ	INJECTIONS	INTRAVENOUS & OTHE
Sale	12018071649	Specialisation4	Department1	8/31/2022	1	0	258.86	319.8	0	Form1	PEPTIDE BASED DIET POWDER	NUTRITIONAL SUPPLEMENTS	NUTRITION

Query:

select distinct Dept from projectfinaldata;

Output:

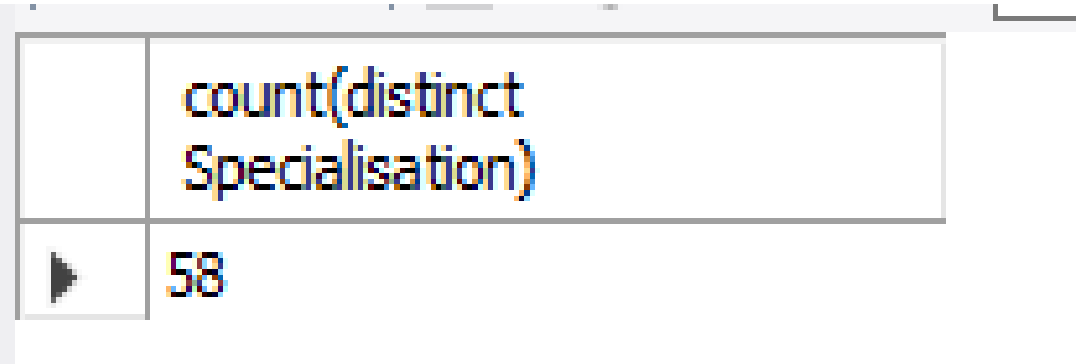


Dept
Department1
Department3
Department2

Query:

select count(distinct Specialisation) from projectfinaldata;

Output:



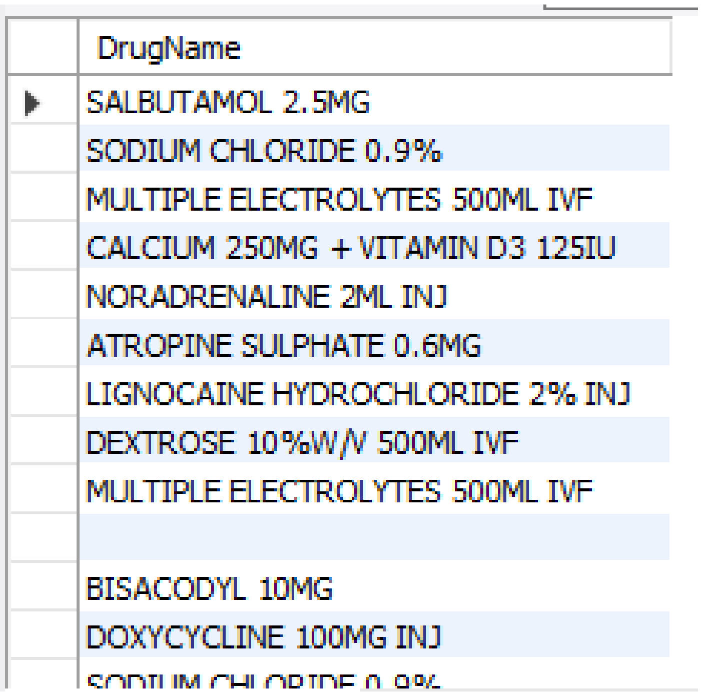
count(distinct Specialisation)
58

Query:

select DrugName from projectfinaldata where Final_Sales=0;

Output:

it was observed that a total of 1638 drugs Final_sales was 0



DrugName
SALBUTAMOL 2.5MG
SODIUM CHLORIDE 0.9%
MULTIPLE ELECTROLYTES 500ML IVF
CALCIUM 250MG + VITAMIN D3 125IU
NORADRENALINE 2ML INJ
ATROPINE SULPHATE 0.6MG
LIGNOCAINE HYDROCHLORIDE 2% INJ
DEXTROSE 10%W/V 500ML IVF
MULTIPLE ELECTROLYTES 500ML IVF
BISACODYL 10MG
DOXYCYCLINE 100MG INJ
SODIUM CHLORIDE 0.9%

So from this insight we can say that these drugs do not contribute much to the sales so these drugs stock intake should be minimized

Query: To handle Null values

```
select count(*),Formulation,DrugName from projectfinaldata
where Formulation=' ' and DrugName=' ';
```

O/P: 164

```
Delete from projectfinaldata where Formulation='' and
DrugName='';
```

Output: 164 rows deleted

```
select count(*) from projectfinaldata where DrugName='' and
SubCat='' and SubCat1='';
```

O/P: 1504

So these 1504 rows can be deleted as we cant get meaningful insight from these columns

```
select count(*) from projectfinaldata where DrugName='' and
SubCat='' and SubCat1='';
```

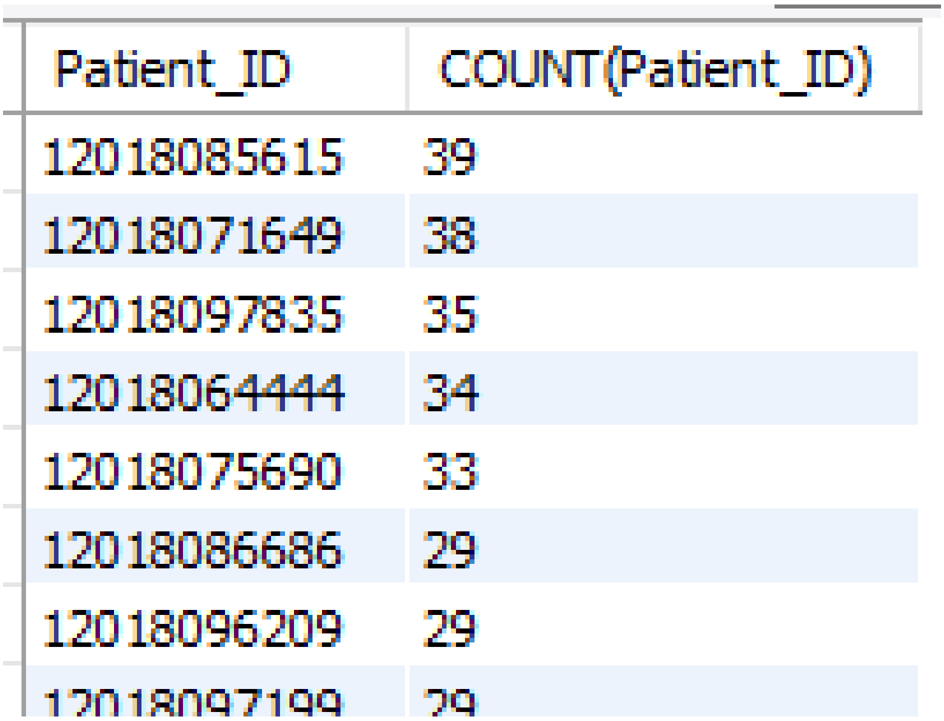
Output: 1504 rows deleted

Query:

```
Select Patient_ID, COUNT(Patient_ID)

from projectfinaldata

group  by Patient_ID

having COUNT(Patient_ID) > 1 order by count(Patient_ID)
desc;
```

This Query selects the frequency of a particular patient in descending order

Output:



Patient_ID	COUNT(Patient_ID)
12018085615	39
12018071649	38
12018097835	35
12018064444	34
12018075690	33
12018086686	29
12018096209	29
12018097199	29

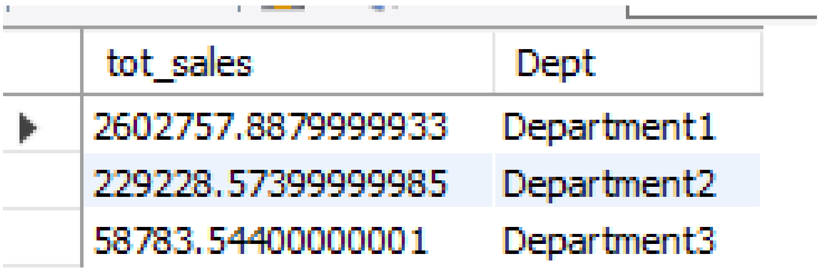
2442 row(s) returned

Query:

select sum(Final_Sales) as tot_sales,Dept from projectfinaldata group by Dept order by tot_sales desc;

This will return which department has the highest sales

Output:



tot_sales	Dept
2602757.8879999933	Department1
229228.57399999985	Department2
58783.54400000001	Department3

From the o/p Department 1 is having the Highest sales of 2602757

```
select * from projectfinaldata where Typeofsales='Return'
order by ReturnQuantity desc;
```

o/p This will return all the rows which has sales type as ‘Returm’

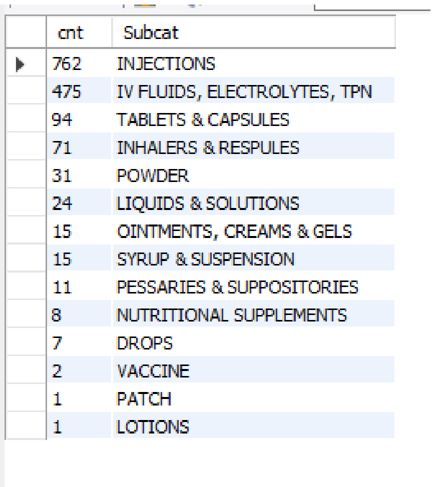
1514 row(s) returned

Query:

```
select count(*) as cnt, Subcat

from (select Subcat from projectfinaldata where
Typeofsales='Return') as sub_table

group by SubCat order by cnt desc;
```

this query will return which subcategory medicines where most frequently returned by the customers/patients

Output:



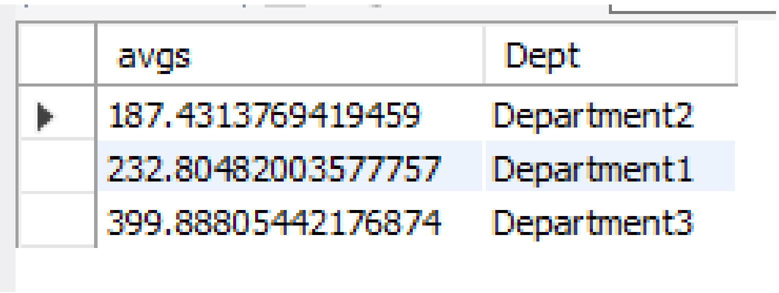
cnt	Subcat
762	INJECTIONS
475	IV FLUIDS, ELECTROLYTES, TPN
94	TABLETS & CAPSULES
71	INHALERS & RESPULES
31	POWDER
24	LIQUIDS & SOLUTIONS
15	OINTMENTS, CREAMS & GELS
15	SYRUP & SUSPENSION
11	PESSARIES & SUPPOSITORIES
8	NUTRITIONAL SUPPLEMENTS
7	DROPS
2	VACCINE
1	PATCH
1	LOTIONS

From the output we can say that Injections are most frequently returned by the patients so we can recommend alternative medicines instead of Injections to patients

Query:

```
select avg(Final_sales)as avgs,Dept from projectfinaldata group by Dept order by avgs;
```

Output:



	avgs	Dept
▶	187.4313769419459	Department2
	232.80482003577757	Department1
	399.88805442176874	Department3

Query: To set the dateofbill column to datetime datatype instead of text

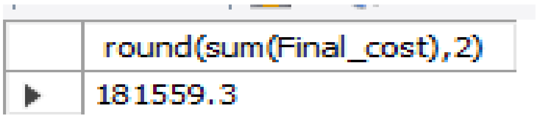
```
update projectfinaldata set Dateofbill=

case

   when Instr(Dateofbill,'-')>0 then str_to_date(Dateofbill,'%d-%m-%Y')

   when Instr(Dateofbill, '/') > 0 then str_to_date(Dateofbill, '%m/%d/%Y')

-- The Instr() function returns an integer value representing the position of the substring within the string. If the substring is found, the function returns the position as a positive integer. If the substring is not found, it returns 0.

end;

alter table projectfinaldata modify column Dateofbill datetime;
```

o/p: 12250 rows affected and column dtype change to datetime

Query:

```
select round(sum(Final_cost),2) from projectfinaldata where
Final_Sales=0;
```

Output:



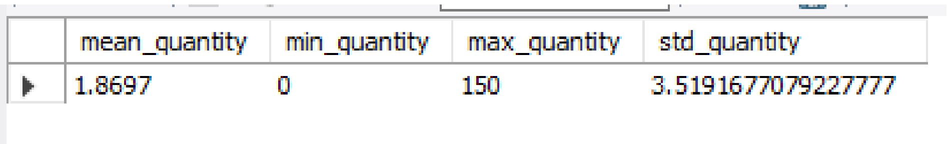
	round(sum(Final_cost),2)
▶	181559.3

Query:

-- Univariate analysis

```
-- Mean, Median, Min, Max, and Std Deviation of Quantity
select
  avg(quantity) as mean_quantity,
  min(quantity) as min_quantity,
  max(quantity) as max_quantity,
  stddev(quantity) as std_quantity
from projectfinaldata;
```

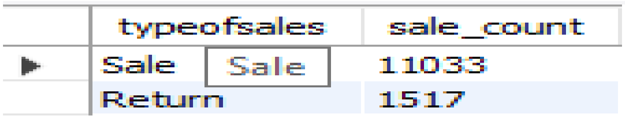
Output:



	mean_quantity	min_quantity	max_quantity	std_quantity
▶	1.8697	0	150	3.519167707922777

```
-- Count occurrences of each type of sale
select typeofsales, count(*) as sale_count
from projectfinaldata group by typeofsales;
```

o/p:

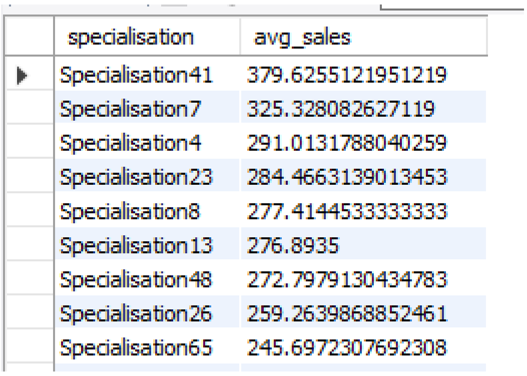


typeofsales	sale_count
Sale	11033
Return	1517

--- Bivariate Analysis

-- Average Final Sales for each Specialisation

```
select specialisation, avg(final_sales) as avg_sales
from projectfinaldata
group by specialisation order by avg_sales desc;
```



specialisation	avg_sales
Specialisation41	379.6255121951219
Specialisation7	325.328082627119
Specialisation4	291.0131788040259
Specialisation23	284.4663139013453
Specialisation8	277.4144533333333
Specialisation13	276.8935
Specialisation48	272.7979130434783
Specialisation26	259.2639868852461
Specialisation65	245.6972307692308

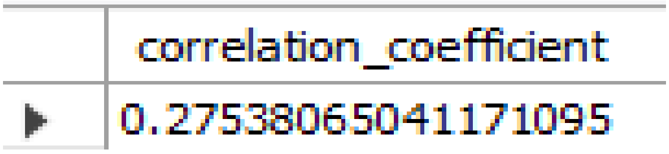
56 row(s) returned

-- Multivariate Analysis

-- correlation between quantity and Final_sales

```
select
  (SUM((Q - mean_quantity) * (F - mean_final_sales)) / (SQRT(SUM((Q - mean_quantity) * (Q - mean_quantity))) * SQRT(SUM((F - mean_final_sales) * (F - mean_final_sales))))) as correlation_coefficient
from (
  select
    Quantity as Q, Final_Sales as F,
    (select avg(Quantity) from projectfinaldata) as mean_quantity,
    (select avg(Final_Sales) from projectfinaldata) as mean_final_sales
  from projectfinaldata) as sub_query;
```

O/P:

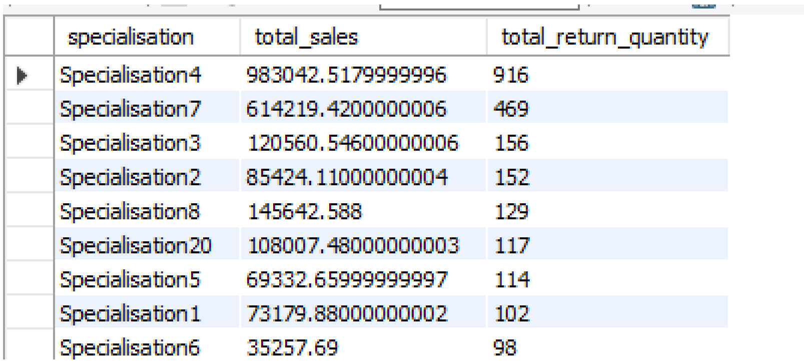


correlation_coefficient
0.2753806504117109‍5

From the result quantity and final_sales are positively correlated i.e as quantity increases the final_sales also increases

```
-- Pivot table to show Total Sales and Total Return Quantity by Specialisation
select specialisation,
      sum(final_sales) as total_sales,
      sum(returnquantity) as total_return_quantity
from projectfinaldata
group by specialisation order by total_return_quantity desc,total_sales;
```

O/P:

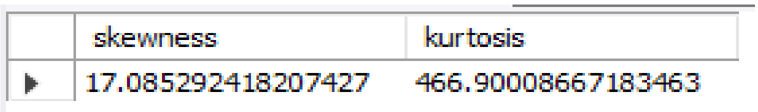


specialisation	total_sales	total_return_quantity
Specialisation4	983042.5179999996	916
Specialisation7	614219.4200000006	469
Specialisation3	120560.54600000006	156
Specialisation2	85424.11000000004	152
Specialisation8	145642.588	129
Specialisation20	108007.48000000003	117
Specialisation5	69332.65999999997	114
Specialisation1	73179.88000000002	102
Specialisation6	35257.69	98

56 row(s) returned

```
-- Skewness and kurtosis
select
  (SUM(POW(Quantity - mean_quantity, 3)) / (COUNT(Quantity) * POW(STDDEV(Quantity), 3))) as skewness,
  (SUM(POW(Quantity - mean_quantity, 4)) / (COUNT(Quantity) * POW(STDDEV(Quantity), 4))) as kurtosis
from projectfinaldata,
  (select avg(Quantity) as mean_quantity from projectfinaldata) as subquery;
```

O/P:



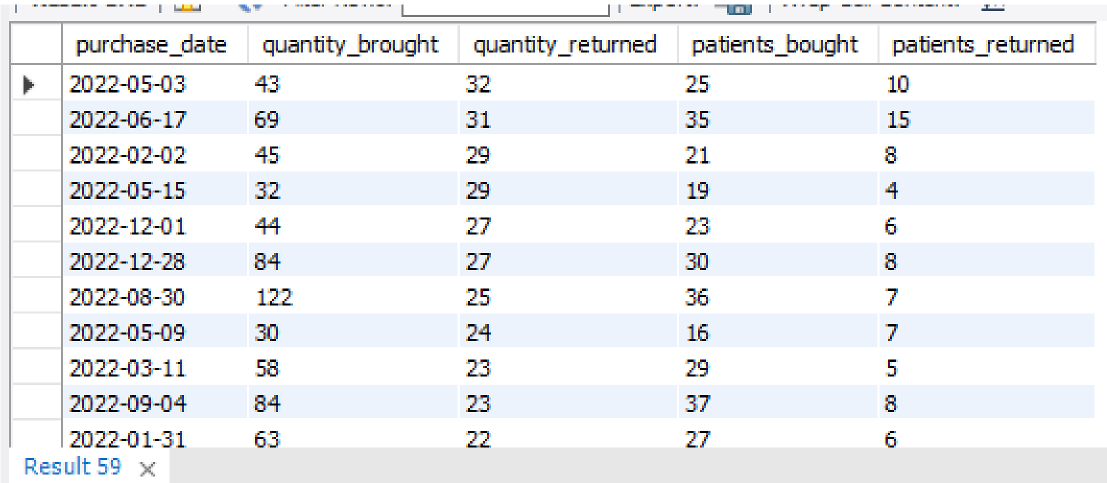
skewness	kurtosis
17.085292418207427	466.90008667183463

A skewness value of 17.08 indicates a highly skewed distribution

it suggests that the distribution of the Quantity column is highly skewed towards higher values.

the kurtosis value of 466.9 indicates that the Quantity column has a distribution with heavy tails and a significant number of outliers.

```
select
  Dateofbill as purchase_date,
  sum(Quantity) as quantity_brought,
  sum(ReturnQuantity) as quantity_returned,
  COUNT(distinct case when Quantity>0 then Patient_ID end) as
patients_bought,
  COUNT(distinct case when ReturnQuantity > 0 then Patient_ID end) as
patients_returned
from projectfinaldata
group by Dateofbill order by quantity_returned desc;
```



purchase_date	quantity_brought	quantity_returned	patients_bought	patients_returned
2022-05-03	43	32	25	10
2022-06-17	69	31	35	15
2022-02-02	45	29	21	8
2022-05-15	32	29	19	4
2022-12-01	44	27	23	6
2022-12-28	84	27	30	8
2022-08-30	122	25	36	7
2022-05-09	30	24	16	7
2022-03-11	58	23	29	5
2022-09-04	84	23	37	8
2022-01-31	63	22	27	6

356 row(s) returned