PHARMA DATA ANALYSIS

Gaurav Singh Rajpoot

INTRODUCTION

During this virtual internship, I explored a Pharma Dataset detailing pharmaceutical sales in Germany and Poland. With 254,083 rows and 18 columns, the dataset formed the backbone of my analysis. Using Microsoft Workbench, I crafted SQL queries to dissect and summarize the sales data. Through this experience, I learned firsthand the importance of working with a clean and validated dataset for meaningful analysis and more evitibly I got a chance to practically work on whatever knowledge I have gained in my academics.. Join me as I share the valuable insights gained from exploring this data.

Q1--Retrieve all columns for all records in the dataset

```
22  Q1--- Retrieve all columns for all records in the dataset.
23
24  Query--- SELECT * FROM pharma;
25
```

Data Output | Messages | Notifications

	distributor character varying	customer character varying	city character varying	country character varying	latitude character varying	longit chara
1	Gottlieb-Cruickshank	Zieme, Doyle and Kunze	Lublin	Poland	51.2333	22.5(
2	Gottlieb-Cruickshank	Feest PLC	Åšwiecie	Poland	53.4167	18.4
3	Gottlieb-Cruickshank	Medhurst-Beer Pharmaceutical Limited	Rybnik	Poland	50.0833	18.5
4	Gottlieb-Cruickshank	Barton Ltd Pharma Plc	CzeladÅ°	Poland	50.3333	19.0
5	Gottlieb-Cruickshank	Keeling LLC Pharmacy	Olsztyn	Poland	53.78	20.4
6	Gottlieb-Cruickshank	Runte-Marquardt Pharmaceutical Ltd	Olecko	Poland	54.0333	22.5
7	Gottlieb-Cruickshank	Blick, Pacocha and Schowalter	InowrocÅ‚aw	Poland	52.7958	18.2(
8	Gottlieb-Cruickshank	Leuschke PLC Pharmacy	CiechanÃ³w	Poland	52.8817	20.6

Q2--How many unique countries are represented in the dataset?

```
31  Q2--- How many unique countries are represented in the dataset?
32
33  Query--- SELECT  DISTINCT country FROM pharma;
34
```

Data Output | Messages | Notifications

	country character varying
1	Germany
2	Poland

Q3--Select the names of all the customers on the 'Retail' channel.

```
51  Q3--- Select the names of all the customers on the 'Retail' channel.
52
53  QUERY--- SELECT customer FROM pharma
54           WHERE sub_channel = 'Retail'
55
```

Data Output | Messages | Notifications

	customer character varying
1	Feest PLC
2	Keeling LLC Pharmacy
3	Blick, Pacocha and Schowalter
4	Leuschke PLC Pharmacy
5	McClure, Zemlak and Dibbert Pharma Plc
6	Lindgren-Simonis Pharm
7	Will and Sons Pharma Plc
8	Jakubowski Inc Pharmaceutical Limited

Q4--Find the total quantity sold for the ' Antibiotics' product class.

```
61  Q4--- Find the total quantity sold for the ' Antibiotics' product class.
62
63  Query--- SELECT sum(quantity) AS Total_Quantity FROM pharma
64           WHERE product_class = 'Antibiotics';
65
```

Data Output | Messages | Notifications

	total_quantity double precision
1	4154321.8570175

Q5-- List all the distinct months present in the dataset

```
70
71  Q5--- List all the distinct months present in the dataset.
72
73  Quantity--- SELECT DISTINCT month FROM pharma;
74
```

Data Output | Messages | Notifications

	month character varying
1	April
2	August
3	December
4	February
5	January
6	July
7	June
8	March

Q6--Calculate the total sales for each year.

```
76  Q6--- Calculate the total sales for each year.
77
78  Query--- SELECT year, SUM(sales) AS Total_Sales FROM pharma
79           GROUP BY year;
80
```

Data Output | Messages | Notifications

	year integer	total_sales double precision
1	2017	2701480740.81559
2	2018	3506897353.6
3	2019	2930937132.7798266
4	2020	2659672415

Q7--Find the customer with the highest sales value.

```
81
82  Q7--- Find the customer with the highest sales value.
83
84  QUERY--- SELECT customer, SUM(sales) AS Total_sales FROM pharma
85           GROUP BY 1
86           ORDER BY 2 DESC
87           LIMIT 1;
88
```

Data Output | Messages | Notifications

	customer character varying	total_sales double precision
1	Mraz-Kutch Pharma Plc	93561780

Q8-- Get the names of all employees who are Sales Reps and are managed by 'James Goodwill'.

```
89  Q8--- Get the names of all employees who are Sales Reps and are managed by 'James Goodwill
90
91  Query--- SELECT DISTINCT name_of_sales_rep AS Employees FROM pharma
92           WHERE manager = 'James Goodwill';
93
```

Data Output | Messages | Notifications

	employees character varying
1	Alan Ray
2	Erica Jones
3	Thompson Crawford

Q9-- Retrieve the top 5 cities with the highest sales.

```
93
94  Q9--- Retrieve the top 5 cities with the highest sales.
95
96  Query--- SELECT city, SUM(sales) AS Total_Sales FROM pharma
97           GROUP BY  1
98           ORDER BY 2 DESC
99           LIMIT 5;
100
```

Data Output | Messages | Notifications

	city character varying	total_sales double precision
1	Butzbach	93561780
2	Baesweiler	64890501
3	Cuxhaven	56006680
4	Friedberg	52183634.6
5	Altenburg	50885320

Q10-- Calculate the average price of products in each sub-channel.

```
102
103  Q10--- Calculate the average price of products in each sub-channel.
104
105  Query--- SELECT sub_channel, AVG(price) AS Average_price FROM pharma
106           GROUP BY 1
107           ORDER BY 2 DESC;
108
```

Data Output | Messages | Notifications

	sub_channel character varying	average_price double precision
1	Government	413.1494398292813
2	Retail	412.80704013108806
3	Institution	411.9543979227524
4	Private	410.71837076539157

Q11-- Join the 'Employees' table with the 'Sales' table to get the name of the Sales Rep and the corresponding sales records.

```
108
109  Q11--- Join the 'Employees' table with the 'Sales' table to get the name of the Sales Rep and the
110         corresponding sales records
111
112  Query--- SELECT name_of_sales_rep, SUM(sales) AS Sales_Records FROM pharma
113           GROUP BY 1
114           ORDER BY 2 DESC;
115
```

Data Output | Messages | Notifications

	name_of_sales_rep character varying	sales_records double precision
1	Jimmy Grey	985969993.944742
2	Abigail Thompson	981056993.864903
3	Sheila Stones	958203898.2441471
4	Daniel Gates	950658635.1859341
5	Anne Wu	920168301.1735809
6	Morris Garcia	901195482.5
7	Stella Given	888340902.41899
8	Jessica Smith	881698369.002429

Q12--Retrieve all sales made by employees from ' Rendsburg ' in the year 2018.

```
115
116  Q12--- Retrieve all sales made by employees from ' Rendsburg ' in the year 2018.
117
118  Query--- SELECT name_of_sales_rep, SUM(sales) AS Total_Sales FROM Pharma
119           WHERE city = 'Rendsburg' AND year = 2018
120           GROUP BY 1
121           ORDER BY 2 DESC;
122
```

Data Output | Messages | Notifications

	name_of_sales_rep character varying	total_sales double precision
1	Jessica Smith	5059318
2	Sheila Stones	1581159
3	Erica Jones	980046
4	Morris Garcia	405500
5	Anne Wu	383869
6	Alan Ray	366832
7	Jimmy Grey	253399
8	Stella Given	226347

Q13-- Calculate the total sales for each product class, for each month, and order the results by year, month, and product class.

```
122
123  Q13--- Calculate the total sales for each product class, for each month, and order
124         the results by year, month, and product class.
125
126  Query--- SELECT product_class, month, year, SUM(sales) AS Total_Sales
127           FROM Pharma
128           GROUP BY 1, 2, 3
129           ORDER BY year, month, product_class;
130
```

Data Output | Messages | Notifications

	product_class character varying	month character varying	year integer	total_sales double precision
1	Analgesics	April	2017	32223716
2	Antibiotics	April	2017	40029226
3	Antimalarial	April	2017	17789675
4	Antipiretics	April	2017	22868812
5	Antiseptics	April	2017	42712211
6	Mood Stabilizers	April	2017	33176944
7	Analgesics	August	2017	49744520
8	Antibiotics	August	2017	32449096

Q14-- Find the top 3 sales reps with the highest sales in 2019.

```
131
132  Q14--- Find the top 3 sales reps with the highest sales in 2019
133
134  Query--- SELECT name_of_sales_rep, SUM(sales) AS Total_Sales
135           FROM Pharma
136           WHERE year = 2019
137           GROUP BY 1
138           ORDER BY 2 DESC
139           LIMIT 3;
140
```

Data Output | Messages | Notifications

	name_of_sales_rep character varying	total_sales double precision
1	Jimmy Grey	310551050.944742
2	Sheila Stones	266924378.244147
3	Daniel Gates	245363929.185934

Q15-- Calculate the monthly total sales for each sub-channel, and then calculate the average monthly sales for each sub-channel over the years.

```
141
142 Q15--- Calculate the monthly total sales for each sub-channel, and then calculate
143       the average monthly sales for each sub-channel over the years.
144
145 Query--- SELECT sub_channel, month, year, SUM(sales) AS Total_Sales,
146          AVG(SUM(sales)) OVER (PARTITION BY sub_channel, month) AS Average_Sales
147          FROM Pharma
148          GROUP BY 1,2,3
149          ORDER BY 3,2;
150
```

Data Output | Messages | Notifications

	sub_channel character varying	month character varying	year integer	total_sales double precision	average_sales double precision
1	Private	April	2017	43680022	38498738.5
2	Retail	April	2017	49076812	53068274.5
3	Institution	April	2017	50151370	49329388.45
4	Government	April	2017	45892380	59112240.75
5	Government	August	2017	61552965	69511663.5
6	Retail	August	2017	67480099	104833982
7	Institution	August	2017	57379276	58881548.75
8	Private	August	2017	47422335	63429645.75

Q16-- Create a summary report that includes the total sales, average price, and total quantity sold for each product class.

```
151
152  Q16--- Create a summary report that includes the total sales, average price, and
153         total quantity sold for each product class.
154
155  Query--- SELECT product_class, SUM(sales) AS Total_Sales,
156           AVG(price) AS Average_Price,
157           SUM(quantity) AS Total_Quantity
158           FROM pharma
159           GROUP BY 1;
160
```

Data Output | Messages | Notifications

	product_class character varying	total_sales double precision	average_price double precision	total_quantity double precision
1	Analgesics	2371515114.283875	432.5710710375187	5553143.783598654
2	Antibiotics	1750277236.54387	419.671056545607	4154321.8570175
3	Antimalarial	1497455333.908921	337.6672080119096	4249075.249670975
4	Antipiretics	1883305591.17649	469.0476796103369	4052544.0572774997
5	Antiseptics	2237524743.6455	412.3966985029883	5499912.71284735
6	Mood Stabilizers	2058909622.636761	400.4933534417753	5169781.142139251

Q17-- Find the top 5 customers with the highest sales for each year.

```
161
162  Q17--- Find the top 5 customers with the highest sales for each year
163
164  Query--- WITH TopCustomers AS (
165           SELECT customer, year, sales,
166           DENSE_RANK() OVER(PARTITION BY year ORDER BY sales DESC) AS Top5
167           FROM Pharma)
168
169           SELECT  * FROM TopCustomers
170           WHERE Top5 <= 5
171
```

Data Output | Messages | Notifications

	customer character varying	year integer	sales double precision	top5 bigint
1	Wiegand, Jast and Yost Pharmaceutical Ltd	2017	17225000	1
2	Fadel-West Pharmaceutical Ltd	2017	14406000	2
3	Kuphal, Herzog and Purdy	2017	13734000	3
4	Abernathy Group Pharmacy	2017	12080000	4
5	Raynor-Graham	2017	10660000	5
6	Watsica, Larson and Labadie Pharmaceutical Ltd	2018	18144000	1
7	Kozey Ltd Pharma Plc	2018	16450000	2
8	Zemlak Group Pharm	2018	16107000	3

Q18-- Calculate the year-over-year growth in sales for each country.

```
171
172  Q18--- Calculate the year-over-year growth in sales for each country.
173
174  Query--- SELECT year, country, SUM(sales) AS Total_Sales,
175           LAG(SUM(sales), 1, 0) OVER(PARTITION BY country ORDER BY year) AS Previous_Year_Sales,
176           SUM(sales) - LAG(SUM(sales), 1, 0) OVER(PARTITION BY country ORDER BY year) AS YearOverYearGrowth
177           FROM pharma
178           GROUP BY 1,2
179           ORDER BY 2,1;
180
```

Data Output | Messages | Notifications

	year integer	country character varying	total_sales double precision	previous_year_sales double precision	yearoveryeargrowth double precision
1	2017	Germany	2701480740.81559	0	2701480740.81559
2	2018	Germany	2826017551.8	2701480740.81559	124536810.98441029
3	2019	Germany	2930937132.779827	2826017551.8	104919580.97982693
4	2020	Germany	2659672415	2930937132.779827	-271264717.7798271
5	2018	Poland	680879801.8	0	680879801.8

Q19--List the months with the lowest sales for each year

```
180
181  Q19--- List the months with the lowest sales for each year
182
183  Query--- WITH Lowest_Sales AS (
184          SELECT month, year, SUM(sales) AS Total_Sales,
185          DENSE_RANK() OVER(PARTITION BY year ORDER BY SUM(sales) ASC) AS Ranks
186          FROM pharma
187          GROUP BY 1,2)
188
189          SELECT * FROM Lowest_Sales
190          WHERE Ranks = 1;
191
```

Data Output | Messages | Notifications

	month character varying	year integer	total_sales double precision	ranks bigint
1	January	2017	151872184	1
2	December	2018	214882167	1
3	January	2019	97664076	1
4	April	2020	135409908	1

Q20-- Calculate the total sales for each sub-channel in each country, and then find the country with the highest total sales for each sub-channel.

```
Q20--- Calculate the total sales for each sub-channel in each country, and then
       find the country with the highest total sales for each sub-channel.

Query--- WITH CTE AS (
         SELECT sub_channel, country, SUM(sales) AS Total_Sales FROM pharma
         GROUP BY 1,2),

         Highest_Total_Sales AS (
         SELECT *, DENSE_RANK() OVER (PARTITION BY sub_channel ORDER BY Total_Sales DESC) AS Ranks
         FROM CTE)

         SELECT * FROM Highest_Total_Sales
         WHERE Ranks = 1;
```

Data Output | Messages | Notifications

	sub_channel character varying	country character varying	total_sales double precision	ranks bigint
1	Government	Germany	2920913380.945978	1
2	Institution	Germany	2719605147.495474	1
3	Private	Germany	2315301981.5627747	1
4	Retail	Germany	3162287330.39119	1

CONCLUSION

In conclusion, the use of SQL queries for data analysis and summarization has yielded insightful information on the pharmaceutical sales environment in Poland and Germany. It was clear from the data analysis which product categories, sub-channels fared the best. The identification of trends in sales volume, average prices, and quantity sold provides practical insights for pricing strategies, inventory control, and market targeting. This practical experience emphasises how crucial it is to work with clean and validated datasets in order to do meaningful analysis and make well-informed decisions in a variety of real-world settings across different sectors.

THANK YOU