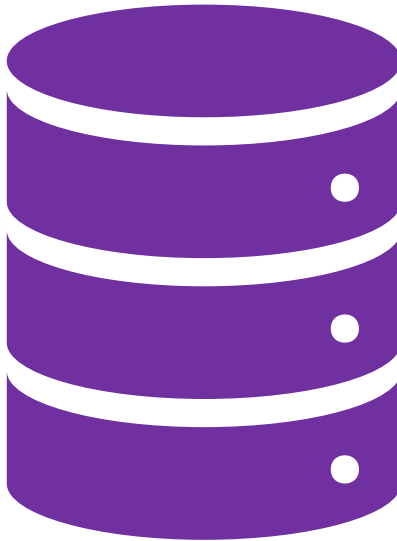# SQL Insights: Analyzing Beverage Store Profits

Ibrahim Najeeb (12/7/2023)

Overview: This SQL project utilizes a Kaggle dataset (link: https://www.kaggle.com/datasets/dsfelix/us-stores-sales). Here's a brief breakdown of the key variables:

Area Code: Unique store identifier.

State: Store location.

Market: Regional market classification.

Market Size: Store size category.

Profit: Profit in dollars ($).

Margin: Profit + Total Expenses ($) or Sales - COGS ($).

Sales: Monetary value from sales.

COGS: Cost of Goods Sold.

Total Expenses: All expenses to bring the product to sale.

Marketing: Marketing-related expenses.

Inventory: Product value in inventory during sale.

Budget Profit: Expected profit in dollars ($).

Budget COGS: Anticipated Cost of Goods Sold ($).

Budget Margin: Expected Profit + Expected Total Expenses ($) or Expected Sales - Expected COGS ($).

Budget Sales: Anticipated value from sales.

ProductID: Unique product identifier.

Date: Sale date.

Product Type: Product category.

Product: Product description.

Type: Product type.

1) **Create a data base for our data set on PGADMIN4**
   ```
   CREATE DATABASE "US_STORE_SALES"
   WITH
   OWNER = postgres
   ENCODING = 'UTF8'
   CONNECTION LIMIT = -1
   IS_TEMPLATE = False;
   ```

2) **Create the table to store our data in**
   ```
   CREATE TABLE public.us_sales
   (
   )
   ;

   ALTER TABLE IF EXISTS public.us_sales
   OWNER to postgres;
   ```

3) **Create columns and insert the value in each column.**
   ```
   ALTER TABLE us_sales
   ADD COLUMN id INTEGER
   ADD COLUMN "Area Code" INTEGER,
   ADD COLUMN State VARCHAR(255),
   ADD COLUMN Market VARCHAR(255),
   ADD COLUMN "Market Size" VARCHAR(255),
   ADD COLUMN Profit INTEGER,
   ADD COLUMN Margin INTEGER,
   ADD COLUMN Sales INTEGER,
   ADD COLUMN COGS INTEGER,
   ADD COLUMN "Total Expenses" INTEGER,
   ADD COLUMN Marketing INTEGER,
   ADD COLUMN Inventory INTEGER,
   ADD COLUMN "Budget Profit" INTEGER,
   ADD COLUMN "Budget COGS" INTEGER,
   ADD COLUMN "Budget Margin" INTEGER,
   ADD COLUMN "Budget Sales" INTEGER,
   ADD COLUMN ProductId INTEGER,
   ADD COLUMN Date,
   ADD COLUMN "Product Type" VARCHAR(255),
   ADD COLUMN Product VARCHAR(255),
   ADD COLUMN Type VARCHAR(255);
   INSERT INTO us_sales (
   ```

id,
"Area Code",
State,
Market,
"Market Size",
Profit,
Margin,
Sales,
COGS,
"Total Expenses",
Marketing,
Inventory,
"Budget Profit",
"Budget COGS",
"Budget Margin",
"Budget Sales",
ProductId,
Date,
"Product Type",
Product,
Type
) **VALUES**
(1,203,'Connecticut','East','Small
Market',107,176,292,116,69,38,962,110,110,160,270,2,'7/1/2010 12:00
AM','Coffee','Columbian','Regular'),
(2,203,'Connecticut','East','Small
Market',75,135,225,90,60,29,1148,90,80,130,210,2,'11/1/201012:00
AM','Coffee','Columbian','Regular'),

4) **Now that we have created our database and table lets inspect the data. Check columns and its data type**

SELECT *
FROM us_sales
LIMIT 5;

	id [PK] integer	Area Code integer	state character varying (255)	market character varying (255)	Market Size character varying (255)	profit integer
1	1	203	Connecticut	East	Small Market	107
2	2	203	Connecticut	East	Small Market	75
3	3	203	Connecticut	East	Small Market	122
4	4	203	Connecticut	East	Small Market	105
5	5	203	Connecticut	East	Small Market	104

5) **Number of product type sold**

SELECT "Product Type", COUNT (*) AS "units_sold"
FROM us_sales
GROUP BY "Product Type";

	Product Type character varying (255) 🔒	units_sold bigint 🔒
1	Coffee	1056
2	Herbal Tea	1056
3	Espresso	1176
4	Tea	960

6) **Top 5 most profitable states & their Market Size**

SELECT DISTINCT state, market, SUM(profit) AS profit
FROM us_sales
GROUP BY state, market
ORDER BY profit DESC
LIMIT 5;

	state character varying (255)	market character varying (255)	Market Size character varying (255)	profit bigint
1	California	West	Major Market	31785
2	Illinois	Central	Major Market	30821
3	Iowa	Central	Small Market	22212
4	New York	East	Major Market	20096
5	Colorado	Central	Major Market	17743

7) Sum of each numeric variable

```
SELECT
    SUM (profit) AS total_profit,
    SUM (margin) AS total_margin,
    SUM (sales) AS total_sales,
    SUM (cogs) AS total_cogs,
    SUM ("Total Expenses") AS total_expenses,
    SUM (Marketing) AS total_marketing,
    SUM (inventory) AS total_inventory
FROM
    us_sales;
```

|	total_profit bigint	total_margin bigint	total_sales bigint	total_cogs bigint	total_expenses bigint	total_marketing bigint	total_inventory bigint
1	259543	443038	819811	358672	229662	132474	3183372

8) States with the highest number of average sales.

```
SELECT
    state,
    ROUND(AVG(sales_per_state),0) AS average_sales
FROM (
    SELECT
        state,
        COUNT(*) AS sales_per_state
    FROM
        us_sales
    GROUP BY
        state
) AS subquery
GROUP BY
    state
ORDER BY average_sales DESC
LIMIT 5;
```

	state character varying (255) 🔒	average_sales numeric 🔒
1	Utah	288
2	California	288
3	Colorado	264
4	Nevada	264
5	Oregon	264
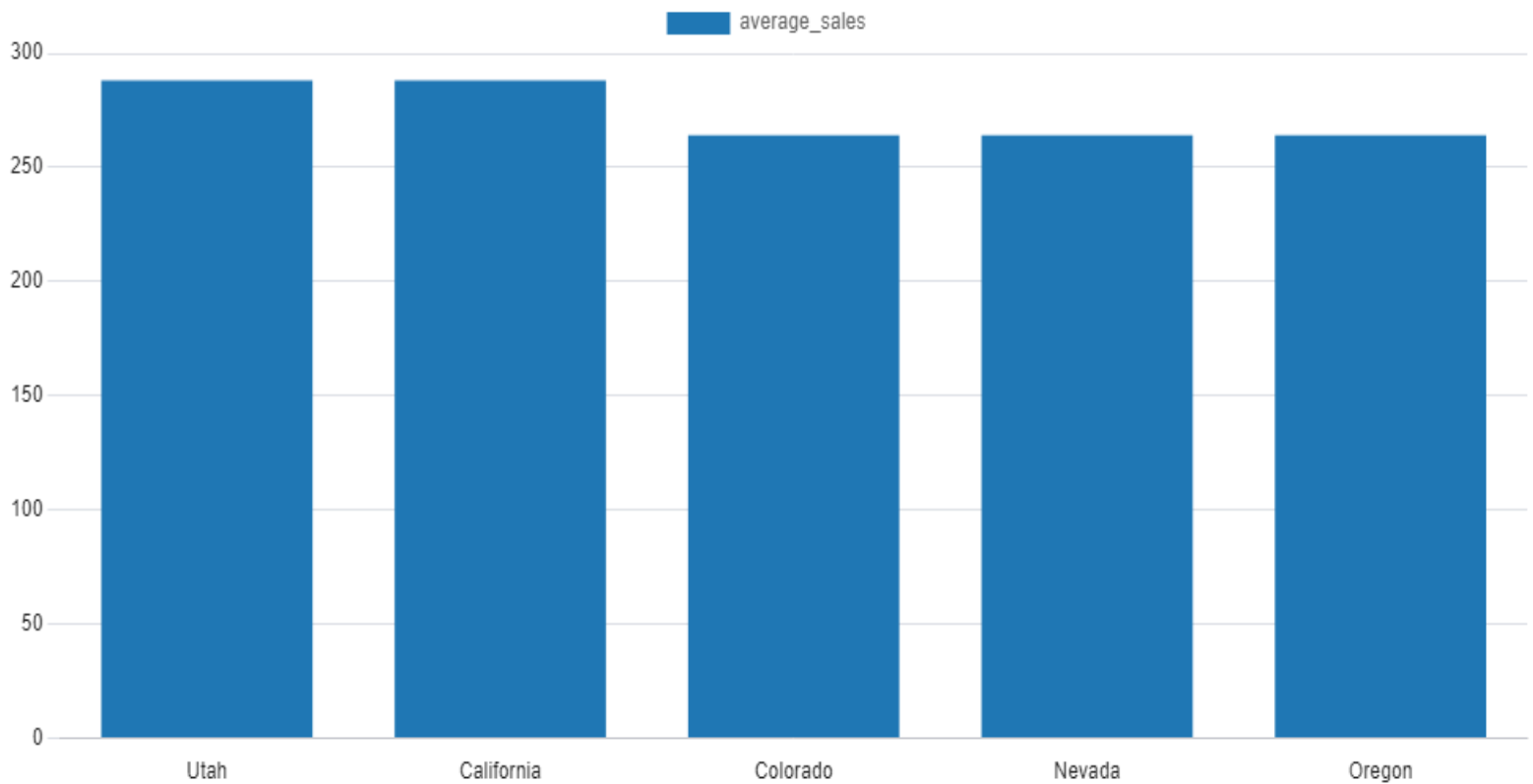
9) **Top 5 states with the most average profit**
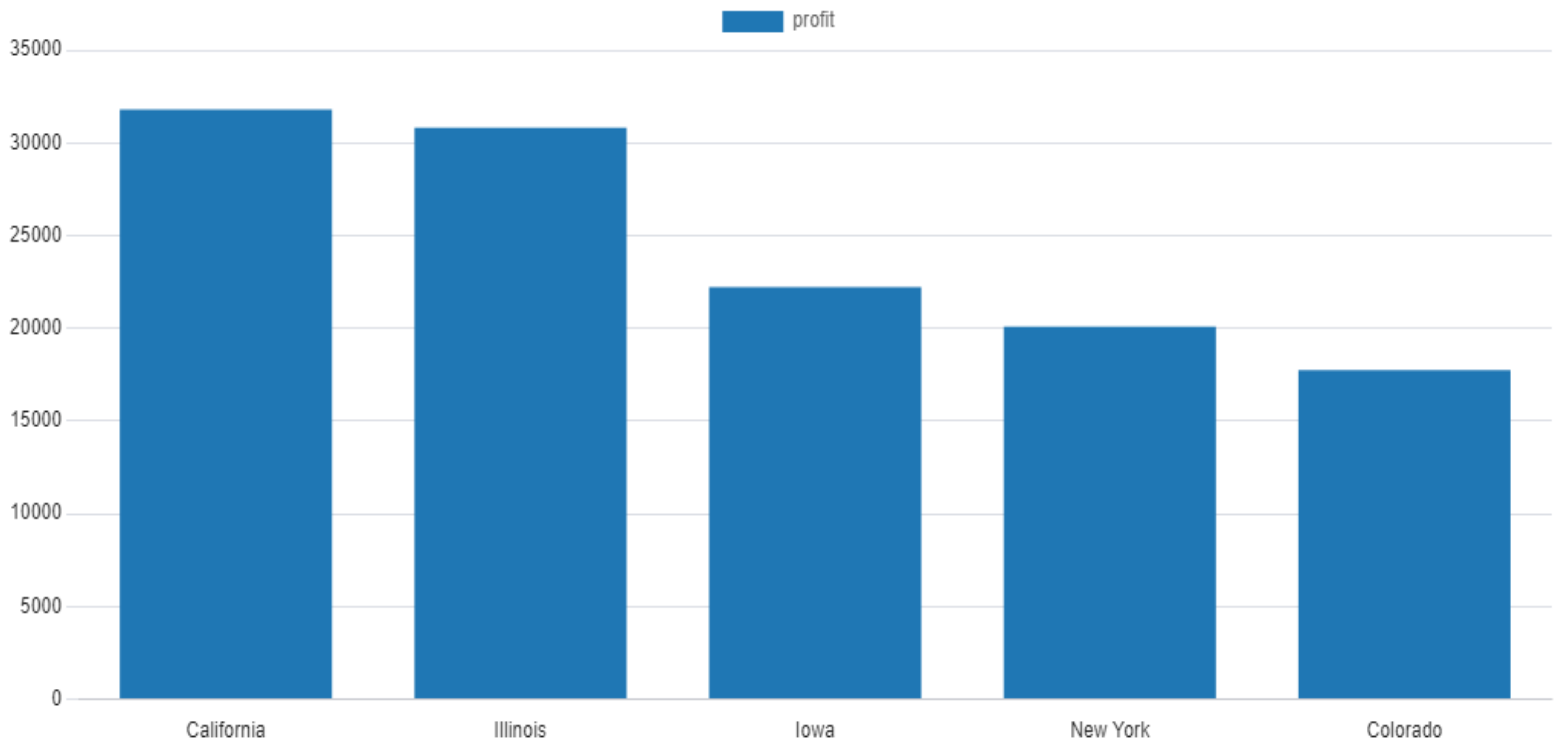
```
SELECT
    state,
    ROUND(AVG(total_profit),0) AS average_profit
FROM (
    SELECT
        state,
        SUM(profit) AS total_profit
    FROM
        us_sales
    GROUP BY
```

```
        state
) AS subquery
GROUP BY
    state
ORDER BY average_profit DESC
LIMIT 5;
```

	state character varying (255) 🔒	average_profit 🔒 numeric
1	California	31785
2	Illinois	30821
3	Iowa	22212
4	New York	20096
5	Colorado	17743

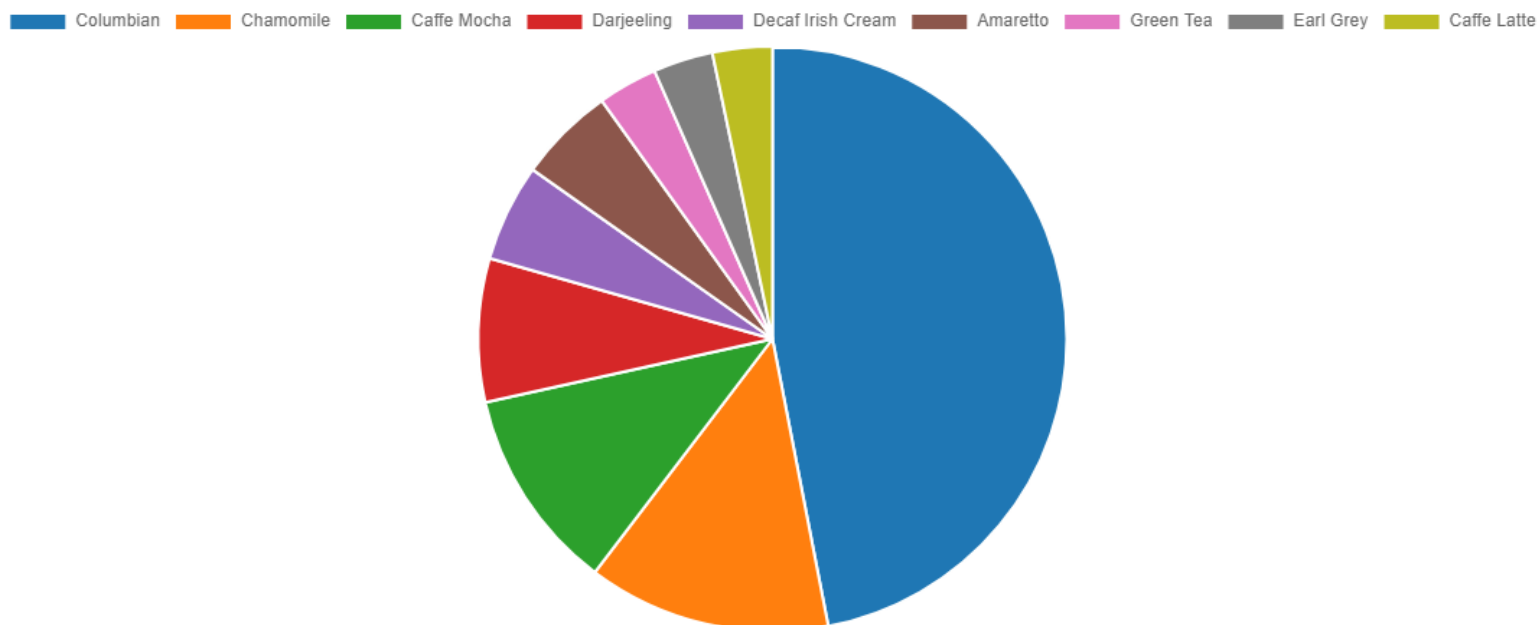10) **Most profitable product in each state**

```sql
WITH ranked_products AS (
  SELECT
    state,
    product,
    SUM(profit) AS total_profit,
    ROW_NUMBER() OVER (PARTITION BY state ORDER BY SUM(profit) DESC) AS product_rank
  FROM
    us_sales
  GROUP BY
    state, product )
SELECT
  state,product,
  total_profit
FROM
  ranked_products
WHERE
  product_rank = 1
ORDER BY
  total_profit DESC;
```

	state character varying (255)	product character varying (255)	total_profit bigint
1	Massachusetts	Columbian	12489
2	California	Columbian	8566
3	New York	Columbian	8565
4	Nevada	Darjeeling	6580
5	Iowa	Chamomile	6577
6	Illinois	Caffe Mocha	6575
7	Texas	Columbian	5452
8	Colorado	Amaretto	3410
9	Connecticut	Columbian	2999
10	Washington	Chamomile	2999
11	Oregon	Green Tea	2726
12	Florida	Decaf Irish Cream	2726
13	Ohio	Earl Grey	2724
14	Oklahoma	Caffe Latte	2723
15	Wisconsin	Decaf Irish Cream	1741
16	Utah	Caffe Mocha	1500
17	Louisiana	Chamomile	1500
18	Missouri	Caffe Mocha	1225
19	New Hampshire	Amaretto	1010
20	New Mexico	Columbian	903

11) **Most profitable products distribution among the states.**
WITH ranked_products AS (
  SELECT
    state,
    product,
    SUM(profit) AS total_profit,
    ROW_NUMBER() OVER (PARTITION BY state ORDER BY SUM(profit) DESC) AS product_rank
  FROM
    us_sales
  GROUP BY
    state, product
), top_products AS(
SELECT
  state,product,
  total_profit
FROM
  ranked_products
WHERE
  product_rank = 1
ORDER BY
  total_profit DESC
)
SELECT DISTINCT product , SUM(total_profit)
FROM top_products
GROUP BY product
ORDER BY SUM DESC

	product character varying (255) 🔒	sum numeric 🔒
1	Columbian	38974
2	Chamomile	11076
3	Caffe Mocha	9300
4	Darjeeling	6580
5	Decaf Irish Cream	4467
6	Amaretto	4420
7	Green Tea	2726
8	Earl Grey	2724
9	Caffe Latte	2723

Legend: Columbian | Chamomile | Caffe Mocha | Darjeeling | Decaf Irish Cream | Amaretto | Green Tea | Earl Grey | Caffe Latte

"Columbian" has a strong dominance overall in the market with no completion, Meanwhile "Chamomile", "Caffe Mocha" & "Darjeeling have a somewhat relevant presence among the states.

12) **Least profitable products in each state**

```
WITH ranked_products AS
(
   SELECT
     state,
     product,
     SUM(profit) AS total_profit,
     ROW_NUMBER() OVER (PARTITION BY state ORDER BY SUM(profit)) AS product_rank
   FROM
     us_sales
   GROUP BY
     state, product
 )
SELECT
   state,product,
   total_profit
FROM
   ranked_products
WHERE
   product_rank = 1
ORDER BY
   total_profit
```

	state character varying (255)	product character varying (255)	total_profit bigint
1	Nevada	Green Tea	-10980
2	New York	Caffe Mocha	-6354
3	California	Decaf Irish Cream	-3891
4	Missouri	Lemon	-686
5	Utah	Mint	-684
6	New Mexico	Decaf Irish Cream	-683
7	Massachusetts	Caffe Mocha	-297
8	New Hampshire	Regular Espresso	-209
9	Colorado	Lemon	-141
10	Oregon	Amaretto	-136
11	Ohio	Decaf Irish Cream	-131
12	Washington	Green Tea	-6
13	Louisiana	Caffe Latte	0
14	Wisconsin	Earl Grey	1
15	Connecticut	Caffe Mocha	3
16	Iowa	Decaf Irish Cream	189
17	Oklahoma	Caffe Mocha	492
18	Florida	Lemon	497
19	Texas	Chamomile	808
20	Illinois	Earl Grey	1368

13) **Least profitable products distribution among the states.**

```sql
WITH ranked_products AS (
  SELECT
    state,
    product,
    SUM(profit) AS total_profit,
    ROW_NUMBER() OVER (PARTITION BY state ORDER BY SUM(profit)) AS product_rank
  FROM
    us_sales
  GROUP BY
    state, product
), top_products AS(
SELECT
  state,product,
  total_profit
FROM
  ranked_products
WHERE
  product_rank = 1
ORDER BY
  total_profit
)
SELECT DISTINCT product , SUM(total_profit)
FROM top_products
GROUP BY product
ORDER BY SUM
```

	product character varying (255) 🔒	sum numeric 🔒
1	Green Tea	-10986
2	Caffe Mocha	-6156
3	Decaf Irish Cream	-4516
4	Mint	-684
5	Lemon	-330
6	Regular Espresso	-209
7	Amaretto	-136
8	Caffe Latte	0
9	Chamomile	808
10	Earl Grey	1369
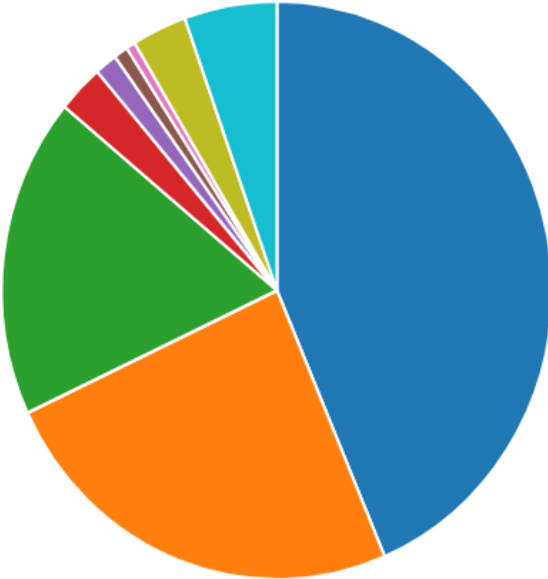
14) **Profit breakdown through months, To see the highest performing month**

SELECT
   SUM(profit) **AS** profit,
   EXTRACT(**MONTH FROM** date::**DATE**) **AS** month
FROM
   us_sales
GROUP BY
   month
ORDER BY
   month ;

	profit 🔒 bigint	month 🔒 numeric
1	23056	1
2	21829	2
3	20258	3
4	20937	4
5	20045	5
6	20761	6
7	21574	7
8	22191	8
9	22094	9
10	22151	10
11	22874	11
12	21773	12
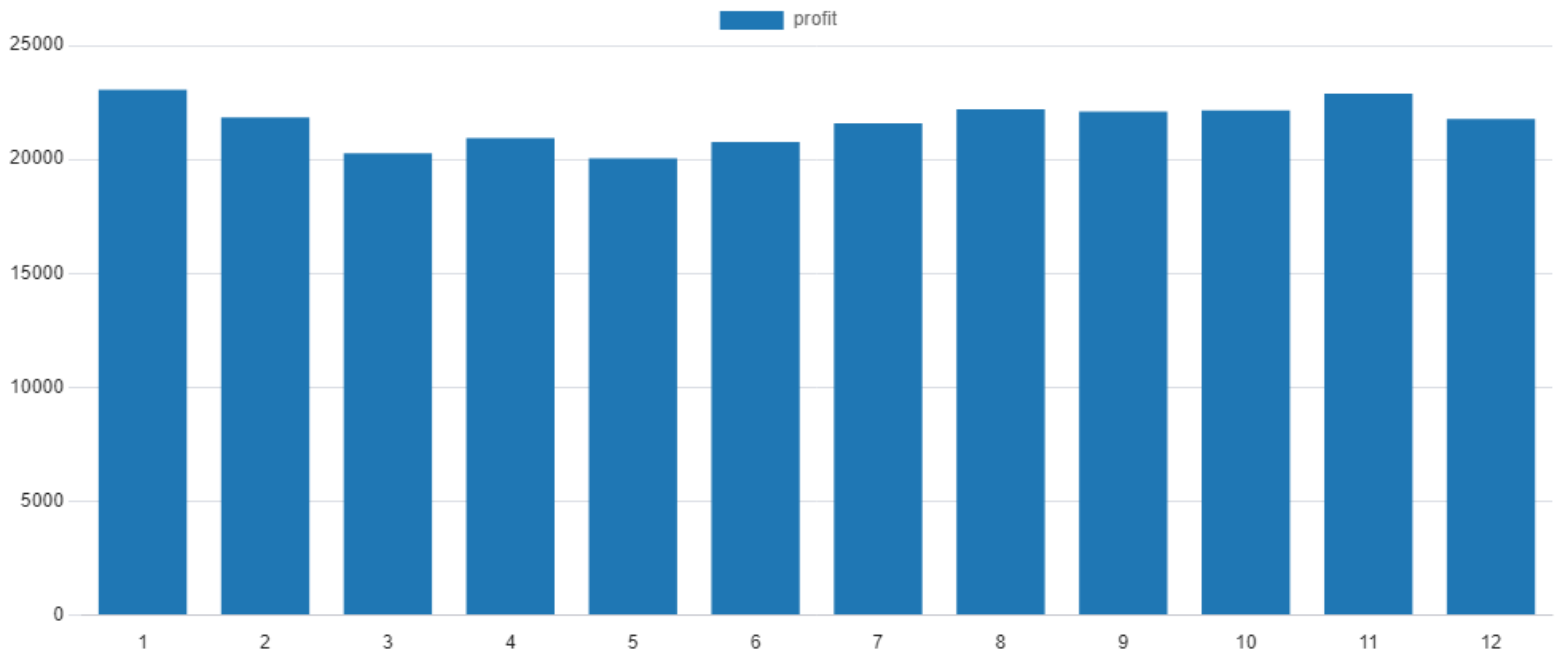
This query calculates the total profit for each month in both 2010 and 2011 combined. However, this approach doesn't provide a clear comparison of how the highest profits in January differ between the two years. To better understand the variation in January profits between 2010 and 2011, we may need to adjust the query to focus specifically on those months in each year.

15) **To compare the profit for the month of January between the years 2010 & 2011.**

```
SELECT
    SUM(CASE WHEN EXTRACT(YEAR FROM date::DATE) = 2010 THEN profit END) AS profit_2010,
    SUM(CASE WHEN EXTRACT(YEAR FROM date::DATE) = 2011 THEN profit END) AS profit_2011,
    (SUM(CASE WHEN EXTRACT(YEAR FROM date::DATE) = 2011 THEN profit END) - SUM(CASE
WHEN EXTRACT(YEAR FROM date::DATE) = 2010 THEN profit END)) AS difference
FROM
    us_sales
WHERE
    EXTRACT(MONTH FROM date::DATE) = 1
```

|	profit_2010 bigint	profit_2011 bigint	difference bigint
1	12405	10651	-1754

Here we can see that collectively, the month of January was the highest profitable month. However, upon closer examination, we observe that in January 2010, the profit was higher compared to January 2011, resulting in a negative difference of approximately 14.1%

16) **With the previous query, we evaluated the performance in 2010 and 2011, focusing on specific months. Now, let's expand the analysis to include product performance between these two years.**

SELECT
  product,
  SUM(CASE WHEN EXTRACT(YEAR FROM date::**DATE)** = 2010 **THEN** profit **END) AS** profit_2010,
  SUM(CASE WHEN EXTRACT(YEAR FROM date::**DATE)** = 2011 **THEN** profit **END) AS** profit_2011,
  (SUM(CASE WHEN EXTRACT(YEAR FROM date::**DATE)** = 2011 **THEN** profit **END)** - **SUM(CASE WHEN EXTRACT(YEAR FROM** date::**DATE)** = 2010 **THEN** profit **END)) AS** difference
FROM
  us_sales
GROUP BY
  product
ORDER BY difference **DESC**;

	product character varying (255)	profit_2010 bigint	profit_2011 bigint	difference bigint
1	Decaf Espresso	13507	15995	2488
2	Earl Grey	10880	13284	2404
3	Columbian	27122	28682	1560
4	Caffe Mocha	8256	9422	1166
5	Lemon	14375	15494	1119
6	Darjeeling	14079	14974	895
7	Regular Espresso	4592	5473	881
8	Amaretto	2287	2603	316
9	Mint	3105	3049	-56
10	Chamomile	13890	13341	-549
11	Decaf Irish Cream	7359	6630	-729
12	Caffe Latte	6054	5321	-733
13	Green Tea	556	-787	-1343
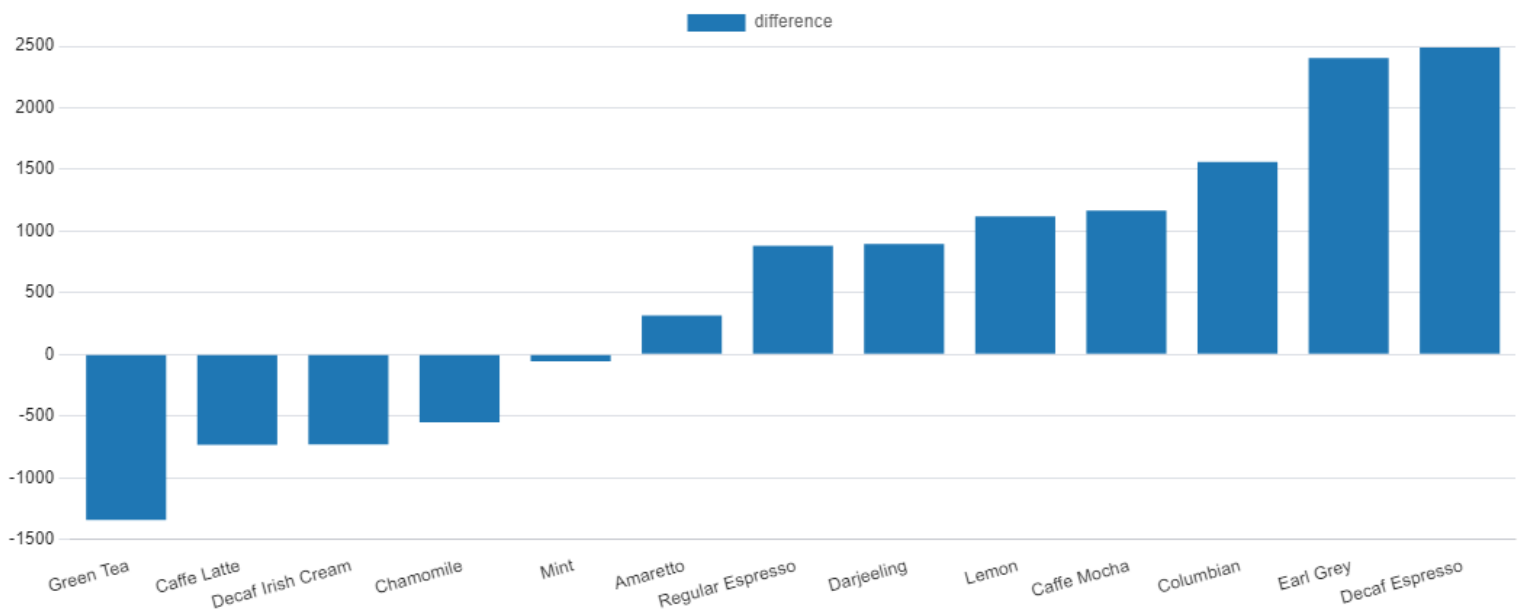
performance of products that either outperformed or underperformed from the previous year to the current year. "Green Tea" is the worst performing product its growth declined from 556 to -787 with a decrease of -1343 compared to "Decaf Espresso" which had a huge growth in terms of profits going from 13507 in 2010 to 15995 in 2011 with a difference of 2488

17) Product Type growth from 2010 to 2011

```
SELECT
    "Product Type",
    SUM(CASE WHEN EXTRACT(YEAR FROM date::DATE) = 2010 THEN profit END) AS profit_2010,
    SUM(CASE WHEN EXTRACT(YEAR FROM date::DATE) = 2011 THEN profit END) AS profit_2011,
    (SUM(CASE WHEN EXTRACT(YEAR FROM date::DATE) = 2011 THEN profit END) -
    SUM(CASE WHEN EXTRACT(YEAR FROM date::DATE) = 2010 THEN profit END)) AS difference
FROM
    us_sales
GROUP BY
    "Product Type"
ORDER BY difference DESC;
```
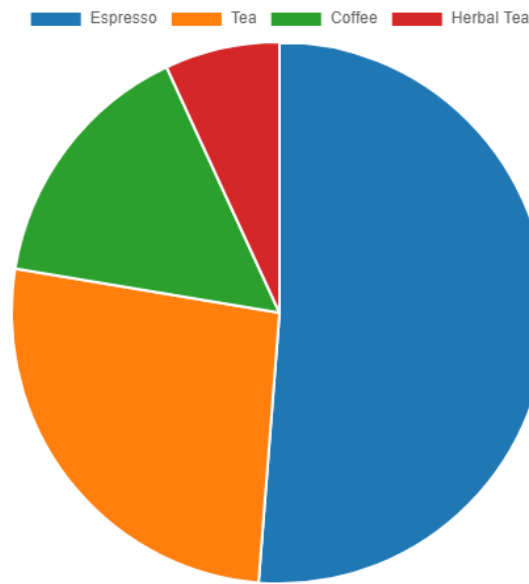
	Product Type character varying (255)	type character varying (255)	profit_2010 bigint	profit_2011 bigint	difference bigint
1	Espresso	Decaf	13507	15995	2488
2	Tea	Regular	25515	27471	1956
3	Coffee	Regular	29409	31285	1876
4	Espresso	Regular	18902	20216	1314
5	Herbal Tea	Decaf	31370	31884	514
6	Coffee	Decaf	7359	6630	-729

Espresso **Tea** **Coffee** **Herbal Tea**

Findings:

- The top 5 most profitable states are California, Illinois, Iowa, New York, and Colorado. Three of these states are in the Central market, and the remaining two are in the West and East markets.
- The total profit for 2010 and 2012 is $259,543.
- Among the top 5 most profitable states, only California is also among the top 5 states with the most average number of sales.
- Higher sales numbers do not necessarily correlate with higher profits for our store.
- "Columbian" dominates the market with no competition. Meanwhile, "Chamomile," "Caffe Mocha," and "Darjeeling" have a relevant presence among the states.
- Tea and herbal tea are the least profitable product types and should be discontinued to avoid further losses. Coffees and Espressos have a strong presence and are generally the most profitable product types.
- In terms of product performance, "Green Tea" is the worst-performing product, experiencing a decline from $556 to -$787, with a decrease of -$1,343. In contrast, "Decaf Espresso" had significant growth, increasing profits from $13,507 in 2010 to $15,995 in 2011, with a difference of $2,488.

Suggestion:

1) Focus on Espresso product types, as they have the highest growth potential overall from 2010 to 2011. There is much potential, and expanding our product line for this type is recommended.
2) Conduct market research on the least performing product type, Tea, to understand the factors limiting its growth, such as demographic considerations.
3) Consider changing the type for Herbal Tea, and for Coffee Decaf, which is underperforming, consider changing the type to Regular to assess its performance.

Analytical Approach:

   i.   Created a PostgreSQL database named "US_STORE_SALES" using PGADMIN4.

  ii.   Defined a table structure for the "us_sales" dataset to store sales data.

 iii.   Inserted sample data into the table for analysis.

 iv.   Explored the dataset by checking the first 5 rows to inspect columns and their data types.

  v.   Conducted product type analysis by counting the number of units sold for each product type.

 vi.   Identified the top 5 most profitable states and their market sizes.

 vii.   Summed up numeric variables for an overall financial overview.

viii.   Determined states with the highest average sales and profit.

 ix.   Analyzed the most profitable product in each state.

  x.   Conducted an analysis of product performance and its distribution among states.

 xi.   Analyzed the total profit breakdown by month.

 xii.   Compared the profit for the month of January between 2010 & 2011.

xiii.   Analyzed the growth of product types from 2010 to 2011.

xiv.   Identified the least profitable products in each state.

 xv.   Analyzed the distribution of least profitable products among states.

xvi.   Summarized findings regarding top profitable states, total profit, and product performance insights.

xvii.   Provided recommendations based on findings for product types and market strategies.