Retail Sales Analysis

1. Project Overview

Analyzing transactional data from 6.6 million customer purchases, this project aims to map market erosion despite high transaction volume thoroughly. The resulting data-driven insights will directly inform **inefficient promotional pricing** and a **strategic roadmap** to optimize pricing models and enhance inventory efficiency.

2. Dataset Summary

- **Total Records:** 6.6 Million transactions
- **Architecture:** Relational model consisting of **7 tables**
- **Columns:** 42 (across all tables)
- **Demographics:** Customer/employee details, including names, gender, and age.
- **Products:** Category, pricing, tiers, and biological attributes.
- **Sales:** Salesperson, customer, product, quantity, full price breakdown, and transaction details.
- **Data Quality**: Minor missing values in name/product attributes; no impact on joins or analysis.

3. Exploratory Data Analysis using Python

We began with data preparation and cleaning in Python:

- **Data Loading:** Imported the dataset using pandas.\
- **Checked for null values**: Used df.isnull().sum() to check for null values in every row.

```
product_id       0
product_name     0
price            0
category_id      0
class            0
modify_date      0
resistant        0
is_allergic      0
vitality_days    0
dtype: int64
```

● **Column Standardization:** Renamed columns to **snake case** for better readability and Documentation.

● **Repeated the above three steps for every table**

● **Feature Engineering:**
  ○ Formatted the **modify_date** column to a datetime object.
  ○ Created **purchase_frequency_days** column from purchase data.
  ○ Rounded up the price column.

● **Sales Table**
  ○ Dropped rows where the sale date is not mentioned using .dropna
  ○ Added **Price** column by merging with the **product table**
  ○ Calculated **Gross Price** and **Net Price**, ordered them accordingly

	sale_id	sales_person_id	customer_id	product_id	quantity	product_price	product_price	gross_price	discount	net_price	sales_date
0	1	6	27039	381	7	44.2	44.2	309.4	0.0	309.40	2018-02-05
1	2	16	25011	61	7	62.6	62.6	438.2	0.0	438.20	2018-02-02
2	3	13	94024	23	24	79.0	79.0	1896.0	0.0	1896.00	2018-05-03
3	4	8	73966	176	19	81.3	81.3	1544.7	0.2	1235.76	2018-04-07
4	5	10	32653	310	9	80.0	80.0	720.0	0.0	720.00	2018-02-12

4. SQL Analysis

Index Creation

```sql
--1. Indexes on Foriegn Keys in the largest table (sales)

CREATE INDEX idx_sales_customer_id ON sales (customer_id);
CREATE INDEX idx_sales_product_id ON sales(product_id);
CREATE INDEX idx_sales_person_id ON sales(sales_person_id);


--2. Index on Time based filtering
CREATE INDEX idx_sales_date ON sales (sales_date);


--3. Indexing Foriegn Keys
CREATE INDEX idx_customer_city_id ON  customer (city_id);
CREATE INDEX idx_employees_city_id ON employees (city_id);
CREATE INDEX idx_product_category_id ON product (category_id);
```

Exploratory Data Analysis (EDA)

- **Category Product Counts** – Shows how many products belong to each category.

	category_name character varying (50)	count bigint
1	Dairy	35
2	Meat	50
3	Seafood	35
4	Poultry	47
5	Grain	28
6	Shell fish	36
7	Cereals	45

- **Product Tier Counts** – Helps identify the distribution of products across tiers.

	product_tier character varying (20)	count bigint
1	High	143
2	Medium	156
3	Low	152

- **Product Price Distribution** – Provides the range and average of product prices.

	min numeric	max numeric	round numeric
1	0.40	99.90	50.91

- **Average Quantity Sold** – Highlights typical purchase quantities.

	round numeric
1	13.00

- **Date Range** – Start to end date of the data

	start_date date	end_date date	duration interval
1	2018-01-01	2018-05-09	4 mons 8 days

- **Quantity Variance & Standard Deviation** – Measures the extent to which quantities fluctuate.

	variance_qty numeric	stdev_qty numeric
1	51.98	7.21

- **Quantity by Discount Status** – Compares purchase behavior with and without discounts.

	discount_cat text	average_quantity_sold numeric
1	Discount	13.01
2	No Discount	13.00

Numerical Outliers

- **Top Discount Outliers** – Identifies unusually high discount values.

	sale_id integer	product_name character varying (100)	discount numeric (4,2)
1	64	Mustard Prepared	0.20
2	66	Lemonade - Natural, 591 Ml	0.20
3	11	Pastry - Raisin Muffin - Mini	0.20
4	13	Sauce - Demi Glace	0.20
5	28	Wine - Chardonnay South	0.20
6	32	Macaroons - Two Bite Choc	0.20
7	8	Bread - Italian Roll With Herbs	0.20
8	4	Smirnoff Green Apple Twist	0.20
9	10	Cheese - Parmesan Grated	0.20
10	71	Cumin - Whole	0.20

- **Monthly Sales Volume** – Shows how sales activity changes month by month.

	sale_year text	sale_month_num text	monthly_sales_count bigint
1	2018	01	1603539
2	2018	02	1448240
3	2018	03	1605559
4	2018	04	1552656
5	2018	05	465813

Financial Performance & Discount Analysis (KPIs)

- **Discount Tier Profitability** – Evaluates revenue generated by different discount levels.

	discount numeric (4,2)	total_transactions bigint	total_revenue numeric	average_order_value_aov numeric
1	0.20	665840	352536666.00	529.46
2	0.10	668823	398970488.16	596.53

- **20% Discount by Category** – Finds which categories respond most to 20% discounts.

	category_id [PK] integer	category_name character varying (50)	total_sales_at_20_percent_discount bigint
1	1	Confections	84046
2	7	Meat	74094
3	9	Poultry	69700
4	3	Cereals	66490
5	11	Produce	63360
6	5	Beverages	56294
7	10	Snails	54686
8	2	Shell fish	52860
9	4	Dairy	51553
10	6	Seafood	51474
11	8	Grain	41283

- **Geographical Revenue Ranking** – Ranks cities based on total revenue.

	city_name character varying (50)	total_transactions bigint	total_net_revenue numeric	revenue_rank bigint
1	Tucson	74736	48348307.27	1
2	Jackson	71623	47919025.53	2
3	Sacramento	73667	47691925.89	3
4	Fort Wayne	74234	47239707.36	4
5	Indianapolis	73637	46919064.01	5
6	Columbus	73993	46862723.39	6
7	Charlotte	73230	46644053.44	7
8	San Antonio	70689	46624164.29	8
9	Phoenix	72988	46361307.83	9
10	Yonkers	72713	46303926.19	10

- **Category Discount AOV** – Measures how discounts affect average order value by category.

	category_name character varying (50)	discount numeric (4,2)	total_transactions bigint	category_aov numeric
1	Beverages	0.20	56294	531.09
2	Beverages	0.10	56722	598.09
3	Cereals	0.20	66490	525.24
4	Cereals	0.10	66456	594.83
5	Confections	0.20	84046	538.52
6	Confections	0.10	84574	605.04
7	Dairy	0.20	51553	554.46
8	Dairy	0.10	51804	627.02
9	Grain	0.20	41283	637.59
10	Grain	0.10	41370	719.18

- **Discount Adoption by City (10%)** – Shows where 10% discounts are used most.

	city_name character varying (50)	total_transactions bigint	ten_percent_discount_transactions bigint	discount_adoption_rate_percent numeric
1	Tucson	74736	7572	10.13
2	San Antonio	70689	7151	10.12
3	Jackson	71623	7233	10.10
4	Charlotte	73230	7389	10.09
5	Sacramento	73667	7427	10.08
6	Indianapolis	73637	7401	10.05
7	Fort Wayne	74234	7322	9.86
8	Columbus	73993	7226	9.77

- **Monthly Average Order Value (AOV)** – Tracks how the average order value shifts across months.

	sales_month timestamp with time zone	total_transactions bigint	monthly_aov numeric
1	2018-01-01 00:00:00+05:30	1603539	642.77
2	2018-02-01 00:00:00+05:30	1448240	641.59
3	2018-03-01 00:00:00+05:30	1605559	642.87
4	2018-04-01 00:00:00+05:30	1552656	642.28
5	2018-05-01 00:00:00+05:30	465813	643.66

Product & Category Health

- **Product Tier Revenue** – Reveals which tiers contribute most to total revenue.

	product_tier character varying (20)	total_transactions bigint	total_units_sold bigint	total_revenue numeric
1	High	2117208	27542974	1448581024.50
2	Low	2249526	29256385	1441564877.74
3	Medium	2309073	30010647	1398971650.52

- **Category Volume Popularity** – Highlights categories with the highest unit sales

	category_name character varying (50)	total_units_sold bigint
1	Confections	10967277
2	Meat	9621629
3	Poultry	9071479
4	Cereals	8647338
5	Produce	8283590
6	Beverages	7320055
7	Snails	7128070
8	Shell fish	6914002
9	Dairy	6745711
10	Seafood	6732271
11	Grain	5378584

- **Product Revenue Ranking** – Lists top products based on total revenue generated.

	product_name character varying (100)	total_revenue numeric	revenue_rank bigint
1	Bread - Calabrese Baguette	18703236.80	1
2	Shrimp - 31/40	18526894.56	2
3	Puree - Passion Fruit	18517036.72	3
4	Tia Maria	18488408.60	4
5	Zucchini - Yellow	18374652.95	5
6	Vanilla Beans	18325803.12	6
7	Grenadine	18155301.20	7
8	Beef - Inside Round	18136797.45	8
9	Lettuce - Treviso	18124887.00	9
10	Tuna - Salad Premix	18024504.66	10

Sales Efficiency & Employee Performance

- **Employee Revenue Leaderboard** – Ranks employees by revenue contribution.

	employee_name text	total_sales_transactions bigint	total_revenue_generated numeric	revenue_rank bigint
1	Devon Brewer	291410	188161280.59	1
2	Shelby Riddle	290017	187576311.06	2
3	Katina Marks	290015	187432169.46	3
4	Desiree Stuart	290068	187266172.88	4
5	Darnell Nielsen	291116	187208747.34	5
6	Julie Dyer	290904	187044863.29	6
7	Wendi Buckley	290286	187034202.92	7
8	Chadwick Cook	290419	186929991.56	8
9	Tonia Mc Millan	289682	186919763.66	9
10	Holly Collins	290448	186882245.97	10

- **High Quantity Transactions** – Identifies employees handling abnormally large orders.

	sales_person_id integer	employee_full_name text	quantity integer
1	12	Lindsay Chen	25
2	13	Katina Marks	25
3	13	Katina Marks	25
4	13	Katina Marks	25
5	10	Jean Vang	25
6	4	Darnell Nielsen	25
7	11	Sonya Dickson	25
8	7	Chadwick Cook	25
9	18	Warren Bartlett	25
10	1	Nicole Fuller	25

- **Employee Transaction Count** – Measures sales activity per employee.

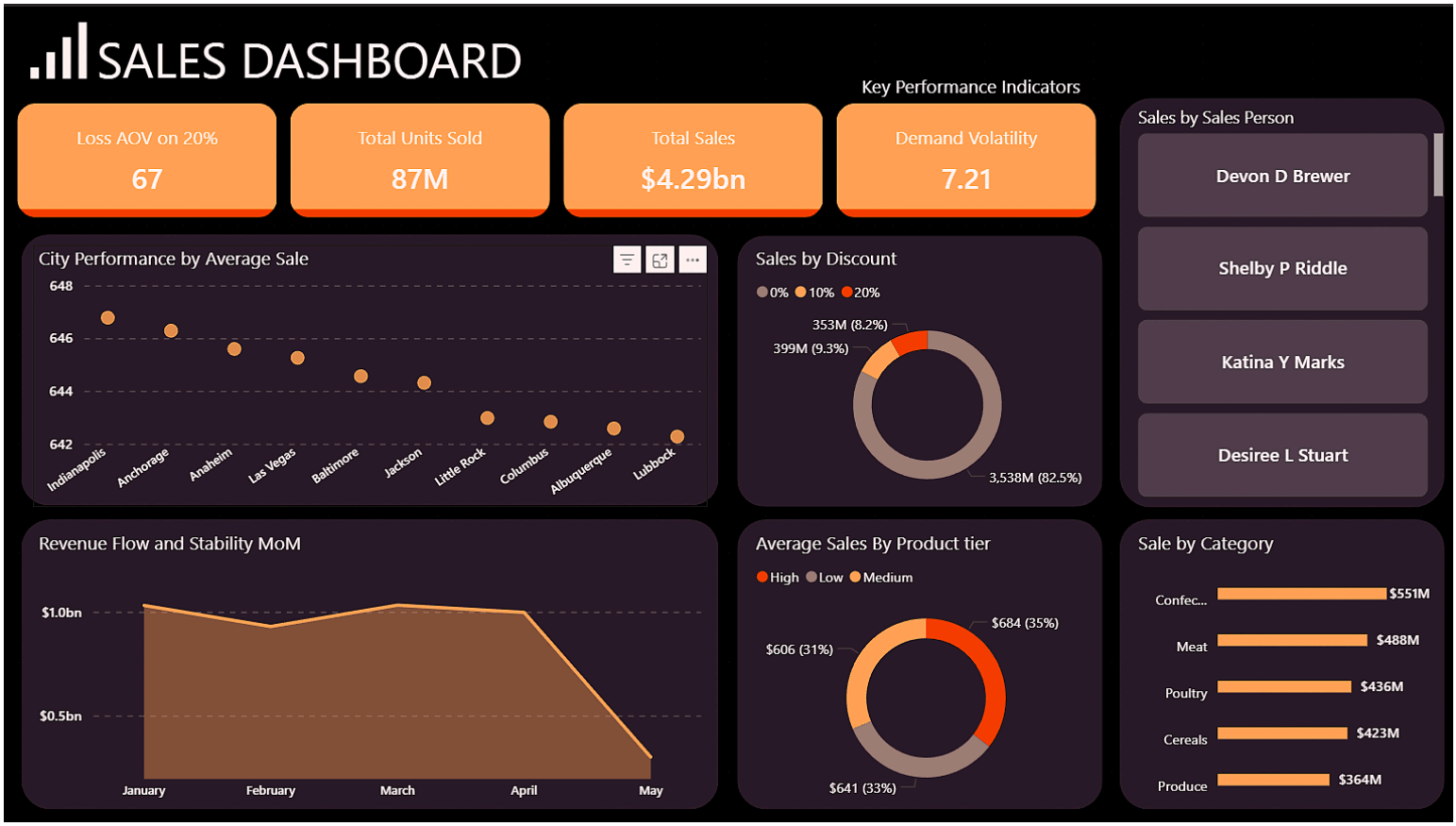
	salesperson_name text	total_sales_transactions bigint
1	Devon Brewer	291410
2	Darnell Nielsen	291116
3	Warren Bartlett	290983
4	Julie Dyer	290904
5	Daphne King	290609
6	Sonya Dickson	290581
7	Holly Collins	290448
8	Chadwick Cook	290419
9	Kari Finley	290393
10	Jean Vang	290373

- **City Quantity per Transaction** – Compares typical order sizes across different cities

	city_name character varying (50)	n_transactions bigint	avg_qty numeric
1	Las Vegas	289682	13.03
2	Anchorage	290015	13.03
3	Baltimore	581858	13.02
4	Fremont	290043	13.01
5	Albuquerque	290232	13.01
6	Anaheim	290068	13.01
7	Indianapolis	290017	13.01
8	Little Rock	290904	13.01
9	Rochester	289693	13.01
10	Tucson	290038	13.00

5. Dashboard in Power BI

Finally, built an interactive dashboard in Power BI to present insights visually.



6. Business Recommendations

- **Eliminate the 20% discount:** Average order value on 20% across all the categories and markets drives lower revenue than 10%.
- **Re-price the Medium Tier:** It has the most volume but generates the lowest revenue, indicating that they are underpriced relative to the demand
- **Maintain Low Safety Stock:** Since the demand is stable, maintain lean inventory to free up capital.
- **Utilize confections as a Traffic Driver:** Confections are the highest volume category; position them strategically to encourage cross-selling of other higher-revenue products.