Data Sources Documentation

1. Introduction:

a. Brief Overview of the Datasets

There are three main datasets in the data warehouse project: Customers_FlipKart, Products_FlipKart, and Sales_FlipKart. The Customers collection keeps track of information about each customer's demographics and behavior, such as what they buy. The Products dataset has important details about products, like their category, price, and availability. The Sales Transactions dataset is the main fact table. It stores information about every purchase and uses unique identifiers to connect customers and goods.

How I created my Dataset?

Using the Faker tool in Jupyter Notebook, I made the datasets for this project. I found the most important fields and relationships for a full retail data warehouse by studying e-commerce data structures on Kaggle. I used a star schema design with Sales as the fact table to make three datasets that are all linked to each other: Customers_FlipKart, Products_FlipKart, and Sales_FlipKart. I made realistic fake data using Faker, such as customer information, product details, and transaction records. Each dataset has more than 1000 records to make sure there is enough information for research.

b. Purpose of the Datasets in the Data Warehouse

These datasets are meant to help people make decisions based on data by making it easy to store, find, and analyze sales and customer contacts. Businesses can learn about the demographics and behavior of their customers with the Customers dataset. The Products dataset, on the other hand, helps them keep track of their supplies, prices, and how well their products are doing. The Sales Transactions dataset is an important part of business intelligence and reporting because it shows trends in income, buying patterns, and operational efficiency.

c. How Datasets are Connected

The three sets of data are linked together using the primary and foreign keys, creating a star schema in the data warehouse. In this case, the Sales Transactions dataset is the fact table, and the customer_id and product_id in the Customers dataset connect them. This structure lets you run quick and effective analytical questions, which lets you keep an eye on what customers buy, how much demand there is for products, and how much money is being made at different levels of detail.

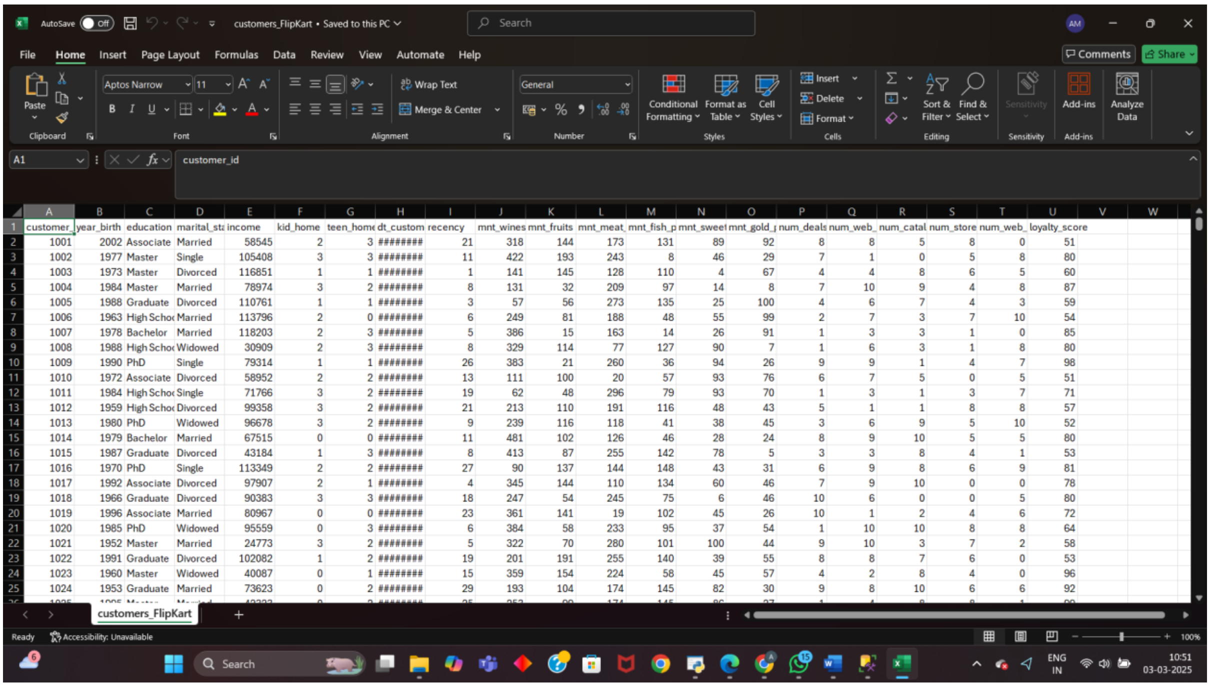
2. Dataset Descriptions:

2.1 Customers_FlipKart Dataset

The Customers dataset has information about customers' demographics and behaviors, which helps businesses look at buying patterns and divide customers into groups. A primary key (customer_id) is used to identify each customer individually. There are fields in the dataset for things like birth year, level of schooling, marital status, income, number of children, purchase history, and loyalty score. For customer segmentation, personalized marketing, and behavior research, this dataset is a must-have.

Key Attributes:

- product_id (Primary Key) – Unique identifier for each product.
- product_name – The name of the product.
- category – The category the product belongs to (e.g., Electronics, Furniture).
- price – The listed price of the product.
- stock_quantity – Available stock levels to monitor inventory.
- rating – Customer review rating for the product.



A1 customer_id

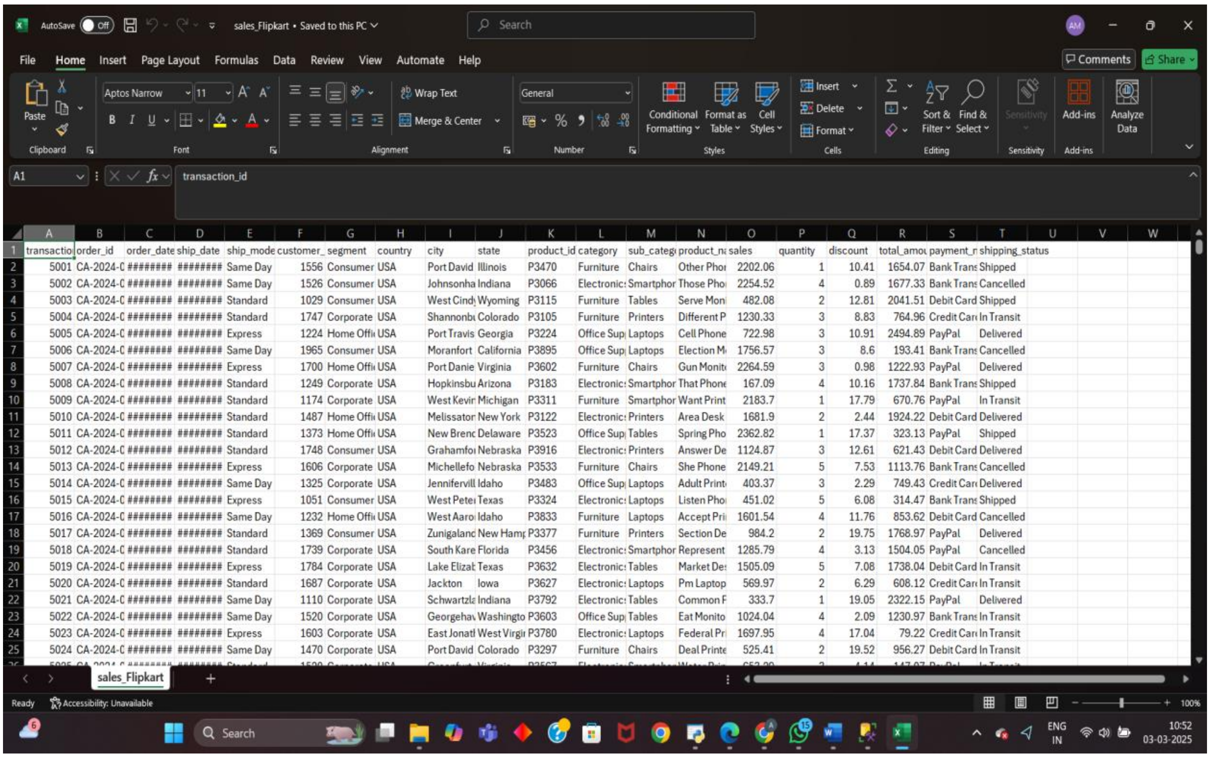
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	customer_	year_birth	education	marital_sta	income	kid_home	teen_home	dt_custom	recency	mnt_wines	mnt_fruits	mnt_meat_	mnt_fish_p	mnt_sweet	mnt_gold_p	num_deals	num_web_	num_catal	num_store	num_web_	loyalty_score
2	1001	2002	Associate	Married	58545	2	3	########	21	318	144	173	131	89	92	8	8	5	8	0	51
3	1002	1977	Master	Single	105408	3	3	########	11	422	193	243	8	46	29	7	1	0	5	8	80
4	1003	1973	Master	Divorced	116851	1	1	########	1	141	145	128	110	4	67	4	4	8	6	5	60
5	1004	1984	Master	Married	78974	3	2	########	8	131	32	209	97	14	8	7	10	9	4	8	87
6	1005	1988	Graduate	Divorced	110761	1	1	########	3	57	56	273	135	25	100	4	6	7	4	3	59
7	1006	1963	High Schoc	Married	113796	2	0	########	6	249	81	188	48	55	99	2	7	3	7	10	54
8	1007	1978	Bachelor	Married	118203	2	3	########	5	386	15	163	14	26	91	1	3	3	1	0	85
9	1008	1988	High Schoc	Widowed	30909	2	3	########	8	329	114	77	127	90	7	1	6	3	1	8	80
10	1009	1990	PhD	Single	79314	1	1	########	26	383	21	260	36	94	26	9	9	1	4	7	98
11	1010	1972	Associate	Divorced	58952	2	2	########	13	111	100	20	57	93	76	6	7	5	0	5	51
12	1011	1984	High Schoc	Single	71766	3	2	########	19	62	48	296	79	93	70	1	3	1	3	7	71
13	1012	1959	High Schoc	Divorced	99358	3	2	########	21	213	110	191	116	48	43	5	1	1	8	8	57
14	1013	1980	PhD	Widowed	96678	3	2	########	9	239	116	118	41	38	45	3	6	9	5	10	52
15	1014	1979	Bachelor	Married	67515	0	0	########	11	481	102	126	46	28	24	8	9	10	5	5	80
16	1015	1987	Graduate	Divorced	43184	1	3	########	8	413	87	255	142	78	5	3	3	8	4	1	53
17	1016	1970	PhD	Single	113349	2	2	########	27	90	137	144	148	43	31	6	9	8	6	9	81
18	1017	1992	Associate	Divorced	97907	2	1	########	4	345	144	110	134	60	46	7	9	10	0	0	78
19	1018	1966	Graduate	Divorced	90383	3	3	########	18	247	54	245	75	6	46	10	6	0	0	5	80
20	1019	1996	Associate	Married	80967	0	0	########	23	361	141	19	102	45	26	10	1	2	4	6	72
21	1020	1985	PhD	Widowed	95559	0	3	########	6	384	58	233	95	37	54	1	10	10	8	8	64
22	1021	1952	Master	Married	24773	3	2	########	5	322	70	280	101	100	44	9	10	3	7	2	58
23	1022	1991	Graduate	Divorced	102082	1	2	########	19	201	191	255	140	39	55	8	8	7	6	0	53
24	1023	1960	Master	Widowed	40087	0	1	########	15	359	154	224	58	45	57	4	2	8	4	0	96
25	1024	1953	Graduate	Married	73623	0	2	########	29	193	104	174	145	82	30	9	8	10	6	6	92

2.2 Sales_FlipKart Dataset

The Sales_FlipKart dataset is the Fact Table in the data center. It keeps track of all the purchases that customers make. It has information about the transaction, like the order details, buy date, customer ID, product ID, quantity, total amount, and payment method. This dataset is very important for tracking sales success, analyzing revenue, and predicting demand because it has foreign keys (customer_id and product_id) that connect customers and products.

Key Attributes:

- transaction_id (Primary Key) – Unique identifier for each transaction.
- customer_id (Foreign Key) – Links to the Customers dataset.
- product_id (Foreign Key) – Links to the Products dataset.
- order_date – Date of the purchase transaction.
- quantity – Number of units purchased in the transaction.
- total_amount – The final purchase amount after discounts.
- payment_method – Payment type used (Credit Card, PayPal, etc.).



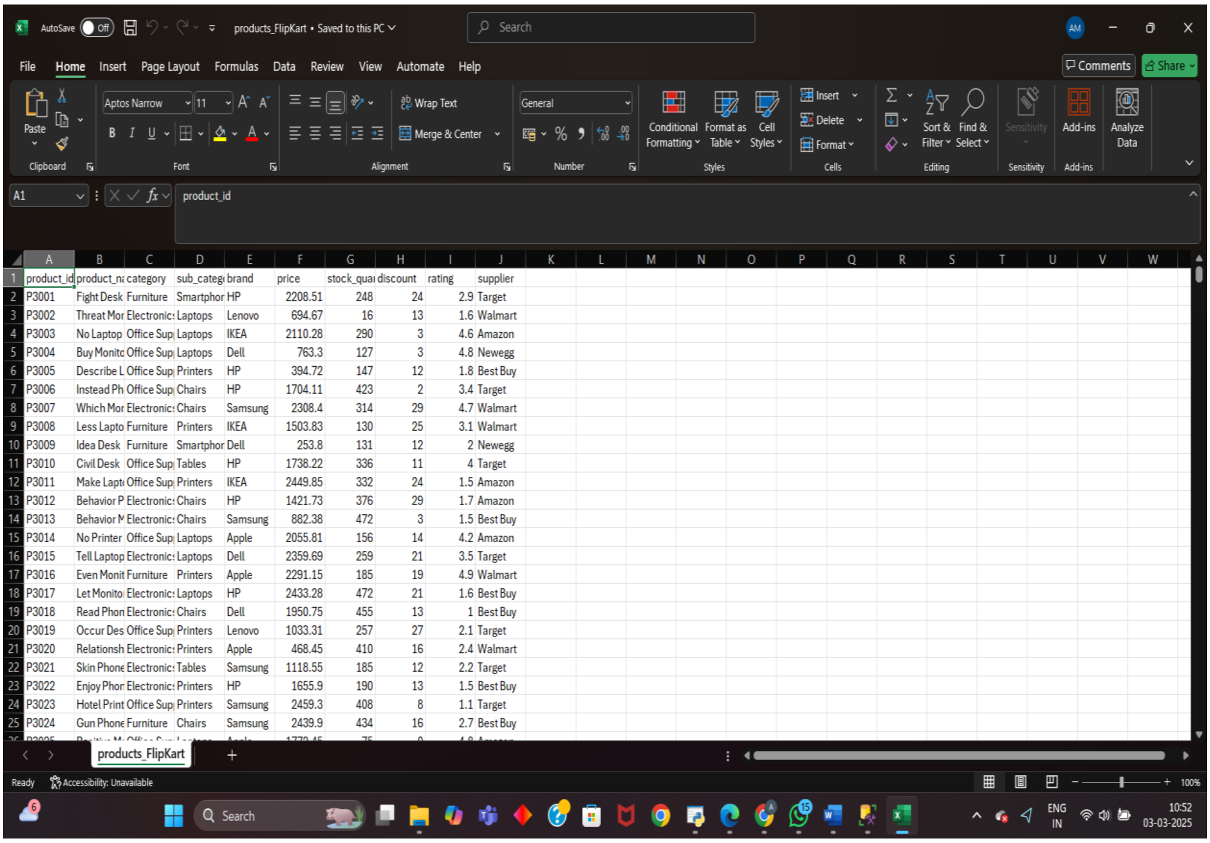
transactio	order_id	order_date	ship_date	ship_mode	customer_	segment	country	city	state	product_id	category	sub_categ	product_na	sales	quantity	discount	total_amou	payment_n	shipping_status
5001	CA-2024-0	########	########	Same Day	1556	Consumer	USA	Port David	Illinois	P3470	Furniture	Chairs	Other Phor	2202.06	1	10.41	1654.07	Bank Trans	Shipped
5002	CA-2024-0	########	########	Same Day	1526	Consumer	USA	Johnsonha	Indiana	P3066	Electronic:	Smartphor	Those Pho	2254.52	4	0.89	1677.33	Bank Trans	Cancelled
5003	CA-2024-0	########	########	Standard	1029	Consumer	USA	West Cindy	Wyoming	P3115	Furniture	Tables	Serve Moni	482.08	2	12.81	2041.51	Debit Card	Shipped
5004	CA-2024-0	########	########	Standard	1747	Corporate	USA	Shannonbu	Colorado	P3105	Furniture	Printers	Different P	1230.33	3	8.83	764.96	Credit Carc	In Transit
5005	CA-2024-0	########	########	Express	1224	Home Offi	USA	Port Travis	Georgia	P3224	Office Sup	Laptops	Cell Phone	722.98	3	10.91	2494.89	PayPal	Delivered
5006	CA-2024-0	########	########	Same Day	1965	Consumer	USA	Moranfort	California	P3895	Office Sup	Laptops	Election M	1756.57	3	8.6	193.41	Bank Trans	Cancelled
5007	CA-2024-0	########	########	Express	1700	Home Offi	USA	Port Danie	Virginia	P3602	Furniture	Chairs	Gun Monit	2264.59	3	0.98	1222.93	PayPal	Delivered
5008	CA-2024-0	########	########	Standard	1249	Corporate	USA	Hopkinsbu	Arizona	P3183	Electronic:	Smartphor	That Phone	167.09	4	10.16	1737.84	Bank Trans	Shipped
5009	CA-2024-0	########	########	Standard	1174	Corporate	USA	West Kevin	Michigan	P3311	Furniture	Smartphor	Want Print	2183.7	1	17.79	670.76	PayPal	In Transit
5010	CA-2024-0	########	########	Standard	1487	Home Offi	USA	Melissator	New York	P3122	Electronic:	Printers	Area Desk	1681.9	2	2.44	1924.22	Debit Card	Delivered
5011	CA-2024-0	########	########	Standard	1373	Home Offi	USA	New Brenc	Delaware	P3523	Office Sup	Tables	Spring Pho	2362.82	1	17.37	323.13	PayPal	Shipped
5012	CA-2024-0	########	########	Standard	1748	Consumer	USA	Grahamfor	Nebraska	P3916	Electronic:	Printers	Answer De	1124.87	3	12.61	621.43	Debit Card	Delivered
5013	CA-2024-0	########	########	Express	1606	Corporate	USA	Michellefo	Nebraska	P3533	Furniture	Chairs	She Phone	2149.21	5	7.53	1113.76	Bank Trans	Cancelled
5014	CA-2024-0	########	########	Same Day	1325	Corporate	USA	Jennifervill	Idaho	P3483	Office Sup	Laptops	Adult Print	403.37	3	2.29	749.43	Credit Carc	Delivered
5015	CA-2024-0	########	########	Express	1051	Consumer	USA	West Peter	Texas	P3324	Electronic:	Laptops	Listen Pho	451.02	5	6.08	314.47	Bank Trans	Shipped
5016	CA-2024-0	########	########	Same Day	1232	Home Offi	USA	West Aaro	Idaho	P3833	Furniture	Laptops	Accept Pri	1601.54	4	11.76	853.62	Debit Card	Cancelled
5017	CA-2024-0	########	########	Standard	1369	Consumer	USA	Zunigaland	New Hamp	P3377	Furniture	Printers	Section De	984.2	2	19.75	1768.97	PayPal	Delivered
5018	CA-2024-0	########	########	Standard	1739	Corporate	USA	South Kare	Florida	P3456	Electronic:	Smartphor	Represent	1285.79	4	3.13	1504.05	PayPal	Cancelled
5019	CA-2024-0	########	########	Express	1784	Corporate	USA	Lake Elizab	Texas	P3632	Electronic:	Tables	Market De:	1505.09	5	7.08	1738.04	Debit Card	In Transit
5020	CA-2024-0	########	########	Standard	1687	Corporate	USA	Jackton	Iowa	P3627	Electronic:	Laptops	Pm Laptop	569.97	2	6.29	608.12	Credit Carc	In Transit
5021	CA-2024-0	########	########	Same Day	1110	Corporate	USA	Schwartzla	Indiana	P3792	Electronic:	Tables	Common F	333.7	1	19.05	2322.15	PayPal	Delivered
5022	CA-2024-0	########	########	Same Day	1520	Corporate	USA	Georgehav	Washingto	P3603	Office Sup	Tables	Eat Monito	1024.04	4	2.09	1230.97	Bank Trans	In Transit
5023	CA-2024-0	########	########	Express	1603	Corporate	USA	East Jonatl	West Virgi	P3780	Electronic:	Laptops	Federal Pr	1697.95	4	17.04	79.22	Credit Carc	In Transit
5024	CA-2024-0	########	########	Same Day	1470	Corporate	USA	Port David	Colorado	P3297	Furniture	Chairs	Deal Printe	525.41	2	19.52	956.27	Debit Card	In Transit

2.3 Products_FlipKart Dataset

The Product_FlipKart dataset has information about items that are for sale, such as their categories, brands, prices, availability, and discounts. A primary key (product_id) is used to identify each product individually. This set of data helps companies keep track of their supplies, study how well their products are doing, and find the best ways to set prices.

Key Attributes:

- product_id (Primary Key) – Unique identifier for each product.
- product_name – The name of the product.
- category – The category the product belongs to (e.g., Electronics, Furniture).
- price – The listed price of the product.
- stock_quantity – Available stock levels to monitor inventory.
- rating – Customer review rating for the product.



product_id	product_na	category	sub_categ	brand	price	stock_quai	discount	rating	supplier
P3001	Fight Desk	Furniture	Smartphor	HP	2208.51	248	24	2.9	Target
P3002	Threat Mor	Electronic:	Laptops	Lenovo	694.67	16	13	1.6	Walmart
P3003	No Laptop	Office Sup	Laptops	IKEA	2110.28	290	3	4.6	Amazon
P3004	Buy Monitc	Office Sup	Laptops	Dell	763.3	127	3	4.8	Newegg
P3005	Describe L	Office Sup	Printers	HP	394.72	147	12	1.8	Best Buy
P3006	Instead Ph	Office Sup	Chairs	HP	1704.11	423	2	3.4	Target
P3007	Which Mor	Electronic:	Chairs	Samsung	2308.4	314	29	4.7	Walmart
P3008	Less Lapto	Furniture	Printers	IKEA	1503.83	130	25	3.1	Walmart
P3009	Idea Desk	Furniture	Smartphor	Dell	253.8	131	12	2	Newegg
P3010	Civil Desk	Office Sup	Tables	HP	1738.22	336	11	4	Target
P3011	Make Lapt	Office Sup	Printers	IKEA	2449.85	332	24	1.5	Amazon
P3012	Behavior P	Electronic:	Chairs	HP	1421.73	376	29	1.7	Amazon
P3013	Behavior M	Electronic:	Chairs	Samsung	882.38	472	3	1.5	Best Buy
P3014	No Printer	Office Sup	Laptops	Apple	2055.81	156	14	4.2	Amazon
P3015	Tell Laptop	Electronic:	Laptops	Dell	2359.69	259	21	3.5	Target
P3016	Even Monit	Furniture	Printers	Apple	2291.15	185	19	4.9	Walmart
P3017	Let Monitor	Electronic:	Laptops	HP	2433.28	472	21	1.6	Best Buy
P3018	Read Phon	Electronic:	Chairs	Dell	1950.75	455	13	1	Best Buy
P3019	Occur Des	Office Sup	Printers	Lenovo	1033.31	257	27	2.1	Target
P3020	Relationsh	Electronic:	Printers	Apple	468.45	410	16	2.4	Walmart
P3021	Skin Phone	Electronic:	Tables	Samsung	1118.55	185	12	2.2	Target
P3022	Enjoy Phor	Electronic:	Printers	HP	1655.9	190	13	1.5	Best Buy
P3023	Hotel Print	Office Sup	Printers	Samsung	2459.3	408	8	1.1	Target
P3024	Gun Phone	Furniture	Chairs	Samsung	2439.9	434	16	2.7	Best Buy

3. Relationships between Datasets:

The Customers, Products, and Sales Transactions datasets are interconnected using primary and foreign keys, forming a star schema model that enables efficient analytical querying.

3.1 Customers_FlipKart Dataset (customer_id)

- The **primary key** customer_id uniquely identifies each customer.
- The **Sales_FlipKart dataset** contains customer_id as a **foreign key**, linking each transaction to a specific customer.

3.2 Products_FlipKart Dataset (product_id)

- The **primary key** product_id uniquely identifies each product.

- The **Sales_FlipKart dataset** contains product_id as a **foreign key**, linking each transaction to the product being purchased.

3.3 Sales_FlipKart Dataset (Fact Table)

- Acts as the **Fact Table**, linking **Customers and Products** datasets.
- Contains both customer_id and product_id as **foreign keys**, ensuring that each sale references a valid customer and product.

Relationship Summary

- **One-to-Many Relationship** (Customers → Sales Transactions):
  - ➢ Each customer can make more than one purchase, but each purchase is only linked to one customer.
- **One-to-Many Relationship** (Products → Sales Transactions):
  - ➢ Every product can be part of more than one transaction, but every transaction is only linked to one product.