Experiment 2

Aim: Data Profiling, Cleaning & Feature Engineering

Objective: Perform profiling, clean issues, engineer new features, validate, and version cleaned dataset.

Detailed Steps

Data Profiling (Pandas Profiling):

Generate a Pandas Profiling Report (now called ydata-profiling) to inspect:

- Data types
- Missing values
- Duplicates
- Outliers
- Value distributions
- Save the report as HTML for reference.

Data Cleaning:

- Handle Missing Values
  - Numeric → replace with median
  - Categorical → replace with mode or "Unknown"
- Remove Duplicates
- Correct Data Types
- Convert columns to correct formats (dates, numeric, categorical).

Feature Engineering:

- Purchase_hour
- Purchase_day_of_week
- Purchase_month
- Is_weekend
- Time_of_day
- Total_amount
- Price_per_rating
- Sentiment_category
- Price_category
- Rating_vs_avg
- High_value_customer

Validation Used:

- We used Great Expectations for comprehensive validation, which included:
- Ensuring the dataset row count is between 5,000 and 15,000
- Checking that required columns product_id and price exist
- Confirming product_id has no missing values
- Verifying rating values fall between 1 and 5
- Verifying price values are between 0 and 1000
- Ensuring customer_gender contains only Male, Female, Other, or Unknown

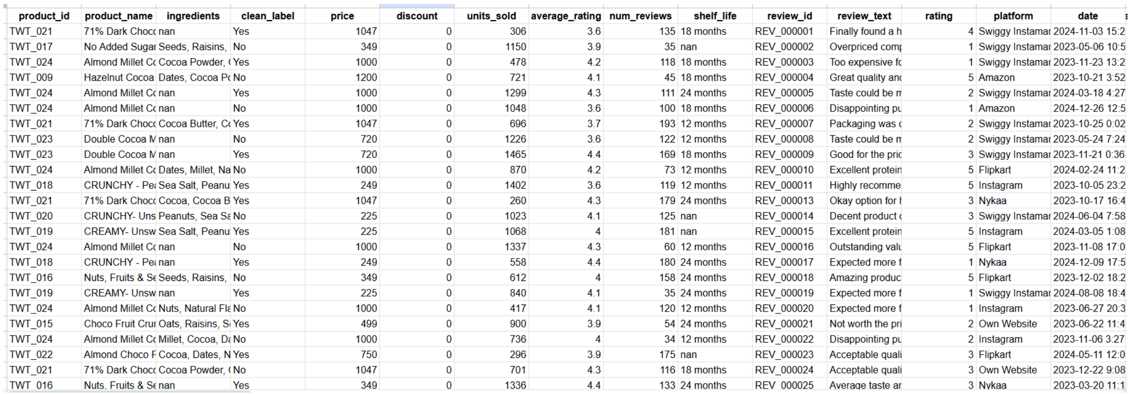
Open-Source Tools Used

Pandas, PyJanitor, Scipy, Pandas Profiling, Great Expectations, DVC

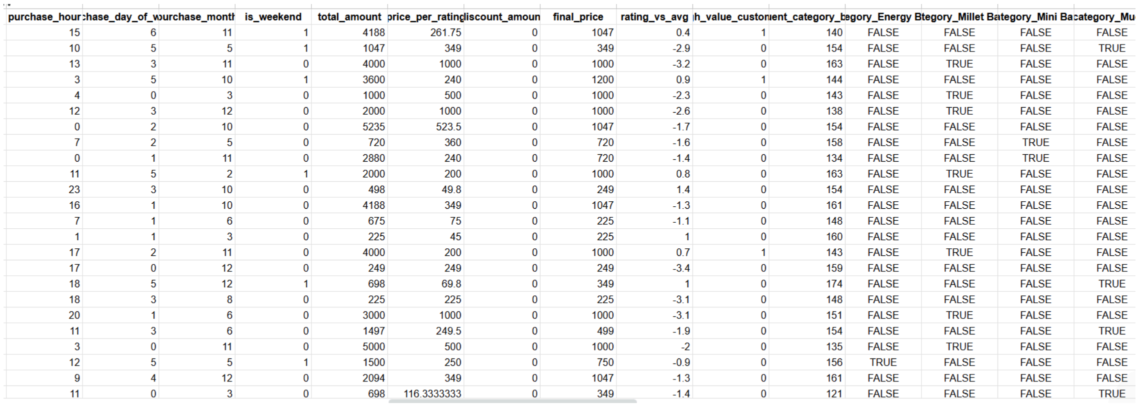
Colab File For Data Cleaning, feature Engineering and Data Profiling (GColab)

Deliverables

- **Unleaned Dataset (GSheet)**
- **Cleaned Dataset (GSheet)**
- **Encoded Cleaned Dataset (GSheet)**

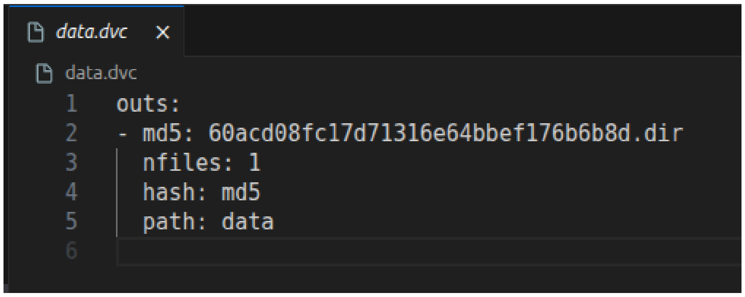


product_id	product_name	ingredients	clean_label	price	discount	units_sold	average_rating	num_reviews	shelf_life	review_id	review_text	rating	platform	date
TWT_021	71% Dark Chocc	nan	Yes	1047	0	306	3.6	135	18 months	REV_000001	Finally found a h	4	Swiggy Instamar	2024-11-03 15:2
TWT_017	No Added Sugar	Seeds, Raisins,	No	349	0	1150	3.9	35	nan	REV_000002	Overpriced comp	1	Swiggy Instamar	2023-05-06 10:5
TWT_024	Almond Millet Cc	Cocoa Powder, (	Yes	1000	0	478	4.2	118	18 months	REV_000003	Too expensive fc	1	Swiggy Instamar	2023-11-23 13:2
TWT_009	Hazelnut Cocoa	Dates, Cocoa Pc	No	1200	0	721	4.1	45	18 months	REV_000004	Great quality anc	5	Amazon	2023-10-21 3:52
TWT_024	Almond Millet Cc	nan	Yes	1000	0	1299	4.3	111	24 months	REV_000005	Taste could be m	2	Swiggy Instamar	2024-03-18 4:27
TWT_024	Almond Millet Cc	nan	No	1000	0	1048	3.6	100	18 months	REV_000006	Disappointing pu	1	Amazon	2024-12-26 12:5
TWT_021	71% Dark Chocc	Cocoa Butter, Cc	Yes	1047	0	696	3.7	193	12 months	REV_000007	Packaging was c	2	Swiggy Instamar	2023-10-25 0:02
TWT_023	Double Cocoa M	nan	No	720	0	1226	3.6	122	12 months	REV_000008	Taste could be m	2	Swiggy Instamar	2023-05-24 7:24
TWT_023	Double Cocoa M	nan	Yes	720	0	1465	4.4	169	18 months	REV_000009	Good for the pric	3	Swiggy Instamar	2023-11-21 0:36
TWT_024	Almond Millet Cc	Dates, Millet, Na	No	1000	0	870	4.2	73	12 months	REV_000010	Excellent protein	5	Flipkart	2024-02-24 11:2
TWT_018	CRUNCHY - Pea	Sea Salt, Peanu	Yes	249	0	1402	3.6	119	12 months	REV_000011	Highly recomme	5	Instagram	2023-10-05 23:2
TWT_021	71% Dark Chocc	Cocoa, Cocoa B	Yes	1047	0	260	4.3	179	24 months	REV_000013	Okay option for l	3	Nykaa	2023-10-17 16:4
TWT_020	CRUNCHY- Uns	Peanuts, Sea Sa	No	225	0	1023	4.1	125	nan	REV_000014	Decent product c	3	Swiggy Instamar	2024-06-04 7:58
TWT_019	CREAMY- Unsw	Sea Salt, Peanu	Yes	225	0	1068	4	181	nan	REV_000015	Excellent protein	5	Instagram	2024-03-05 1:08
TWT_024	Almond Millet Cc	nan	No	1000	0	1337	4.3	60	12 months	REV_000016	Outstanding valu	5	Flipkart	2023-11-08 17:0
TWT_018	CRUNCHY - Pea	nan	Yes	249	0	558	4.4	180	24 months	REV_000017	Expected more f	1	Nykaa	2024-12-09 17:5
TWT_016	Nuts, Fruits & Se	Seeds, Raisins,	No	349	0	612	4	158	24 months	REV_000018	Amazing produc	5	Flipkart	2023-12-02 18:2
TWT_019	CREAMY- Unsw	nan	Yes	225	0	840	4.1	35	24 months	REV_000019	Expected more f	1	Swiggy Instamar	2024-08-08 18:4
TWT_024	Almond Millet Cc	Nuts, Natural Fla	No	1000	0	417	4.1	120	12 months	REV_000020	Expected more f	1	Instagram	2023-06-27 20:3
TWT_015	Choco Fruit Crur	Oats, Raisins, S	Yes	499	0	900	3.9	54	24 months	REV_000021	Not worth the pri	2	Own Website	2023-06-22 11:4
TWT_024	Almond Millet Cc	Millet, Cocoa, Da	No	1000	0	736	4	34	12 months	REV_000022	Disappointing pu	2	Instagram	2023-11-06 3:27
TWT_022	Almond Choco F	Cocoa, Dates, N	Yes	750	0	296	3.9	175	nan	REV_000023	Acceptable quali	3	Flipkart	2024-05-11 12:0
TWT_021	71% Dark Chocc	Cocoa Powder, (	No	1047	0	701	4.3	116	18 months	REV_000024	Acceptable quali	3	Own Website	2023-12-22 9:08
TWT_016	Nuts, Fruits & Se	nan	Yes	349	0	1336	4.4	133	24 months	REV_000025	Average taste ar	3	Nykaa	2023-03-20 11:1



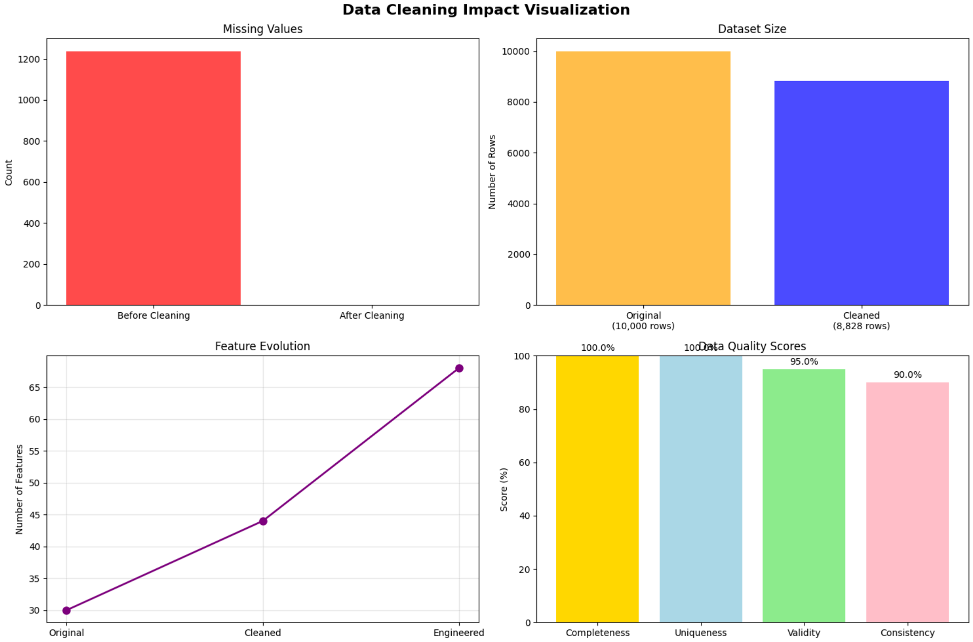
purchase_hour	chase_day_of_w	urchase_montl	is_weekend	total_amount	price_per_rating	discount_amoun	final_price	rating_vs_avg	h_value_custon	ient_category_t	egory_Energy B	tegory_Millet B	ategory_Mini Ba	category_Mu
15	6	11	1	4188	261.75	0	1047	0.4	1	140	FALSE	FALSE	FALSE	FALSE
10	5	5	1	1047	349	0	349	-2.9	0	154	FALSE	FALSE	FALSE	TRUE
13	3	11	0	4000	1000	0	1000	-3.2	0	163	FALSE	TRUE	FALSE	FALSE
3	5	10	1	3600	240	0	1200	0.9	1	144	FALSE	FALSE	FALSE	FALSE
4	0	3	0	1000	500	0	1000	-2.3	0	143	FALSE	TRUE	FALSE	FALSE
12	3	12	0	2000	1000	0	1000	-2.6	0	138	FALSE	TRUE	FALSE	FALSE
0	2	10	0	5235	523.5	0	1047	-1.7	0	154	FALSE	FALSE	FALSE	FALSE
7	2	5	0	720	360	0	720	-1.6	0	158	FALSE	FALSE	TRUE	FALSE
0	1	11	0	2880	240	0	720	-1.4	0	134	FALSE	FALSE	TRUE	FALSE
11	5	2	1	2000	200	0	1000	0.8	0	163	FALSE	TRUE	FALSE	FALSE
23	3	10	0	498	49.8	0	249	1.4	0	154	FALSE	FALSE	FALSE	FALSE
16	1	10	0	4188	349	0	1047	-1.3	0	161	FALSE	FALSE	FALSE	FALSE
7	1	6	0	675	75	0	225	-1.1	0	148	FALSE	FALSE	FALSE	FALSE
1	1	3	0	225	45	0	225	1	0	160	FALSE	FALSE	FALSE	FALSE
17	2	11	0	4000	200	0	1000	0.7	1	143	FALSE	TRUE	FALSE	FALSE
17	0	12	0	249	249	0	249	-3.4	0	159	FALSE	FALSE	FALSE	FALSE
18	5	12	1	698	69.8	0	349	1	0	174	FALSE	FALSE	FALSE	TRUE
18	3	8	0	225	225	0	225	-3.1	0	148	FALSE	FALSE	FALSE	FALSE
20	1	6	0	3000	1000	0	1000	-3.1	0	151	FALSE	TRUE	FALSE	FALSE
11	3	6	0	1497	249.5	0	499	-1.9	0	154	FALSE	FALSE	FALSE	TRUE
3	0	11	0	5000	500	0	1000	-2	0	135	FALSE	TRUE	FALSE	FALSE
12	5	5	1	1500	250	0	750	-0.9	0	156	TRUE	FALSE	FALSE	FALSE
9	4	12	0	2094	349	0	1047	-1.3	0	161	FALSE	FALSE	FALSE	FALSE
11	0	3	0	698	116.3333333	0	349	-1.4	0	121	FALSE	FALSE	FALSE	TRUE

- **Data Profiling (Link)**
- **Data Processing Report (Link)**
- **DVC**



```
outs:
- md5: 60acd08fc17d71316e64bbef176b6b8d.dir
  nfiles: 1
  hash: md5
  path: data
```

- **Visualization**



Conclusion

Through this experiment, I learned how to profile, clean, and enhance a dataset, ensuring accuracy and consistency. I also understood the importance of feature engineering, validation, and version control to maintain quality and reproducibility in data science work. Overall, it strengthened my ability to prepare reliable data for analysis and decision-making.