Marketing and Retail Analytics

By Chayan Banerjee

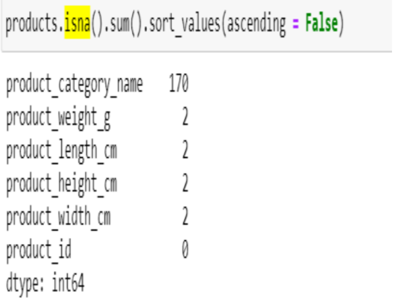
Contents

- Data Cleaning In Python and Exporting the cleaned data set into Tableau
- Analysis using Tableau
- Market basket analysis in Python
- Insights from Market Basket Analysis using association rules
- conclusions

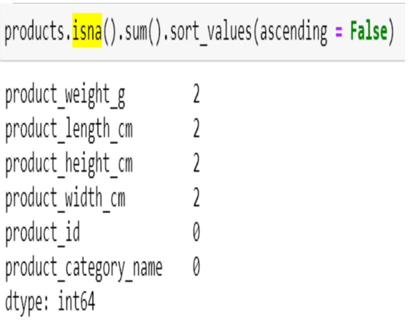
Data Cleaning

- Filtered Data with records containing order status as ‘Delivered’ from Table-Orders
- Missing values Treatment in the columns named in order_approved_at and order_delivered_timestamp
    - Assuming Order_approved_at = Order_purchase_timestamp time is very less
    - Assuming Delivered Timestamp and Estimated delivery date to be same
- Dropping Duplicates in customer_id in Customers Table
- After describing the data, the Mean and the 50% distribution is far away so filled values as per median in columns weight, length, height and width columns in table Products.
        - Weight: 700
        - Length: 25
        - Height: 13
        - Width: 20

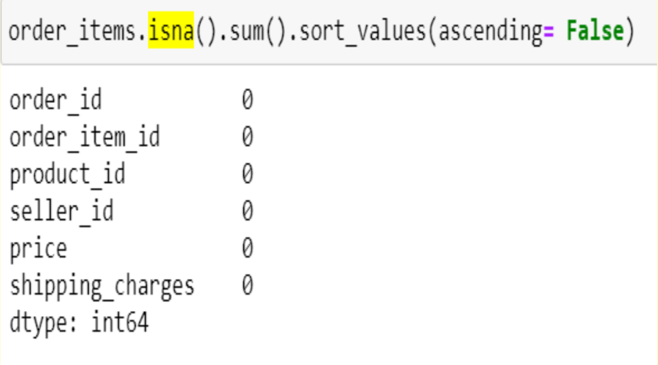
Cleaned data in the sheets as payments, orders, order_items, customers and products



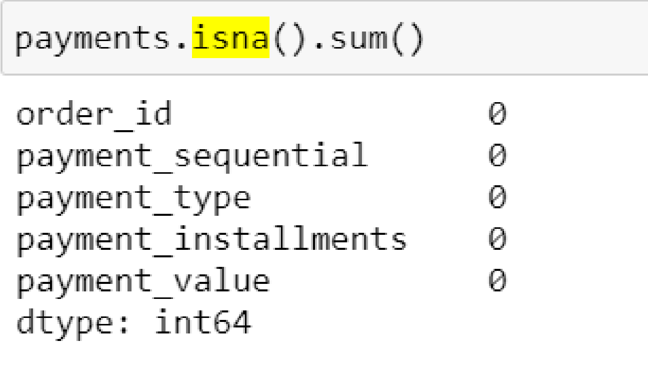
```
products.isna().sum().sort_values(ascending = False)

product_category_name    170
product_weight_g           2
product_length_cm          2
product_height_cm          2
product_width_cm           2
product_id                 0
dtype: int64
```



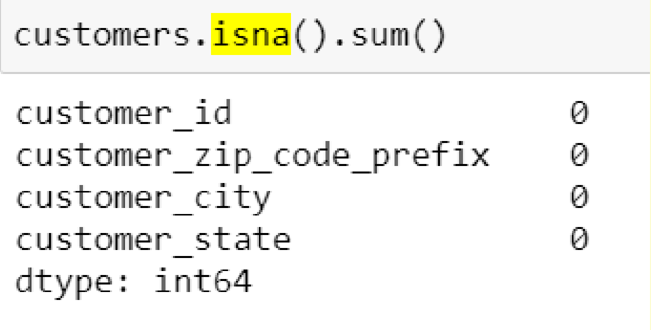
```
products.isna().sum().sort_values(ascending = False)

product_weight_g         2
product_length_cm        2
product_height_cm        2
product_width_cm         2
product_id               0
product_category_name    0
dtype: int64
```



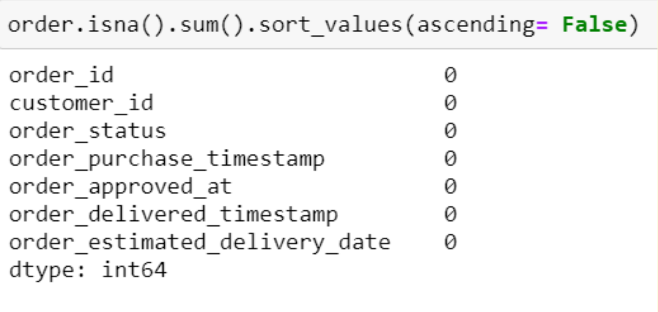
```
order_items.isna().sum().sort_values(ascending= False)

order_id            0
order_item_id       0
product_id          0
seller_id           0
price               0
shipping_charges    0
dtype: int64
```



```
payments.isna().sum()

order_id                0
payment_sequential      0
payment_type            0
payment_installments    0
payment_value           0
dtype: int64
```



```
customers.isna().sum()

customer_id                 0
customer_zip_code_prefix    0
customer_city               0
customer_state              0
dtype: int64
```



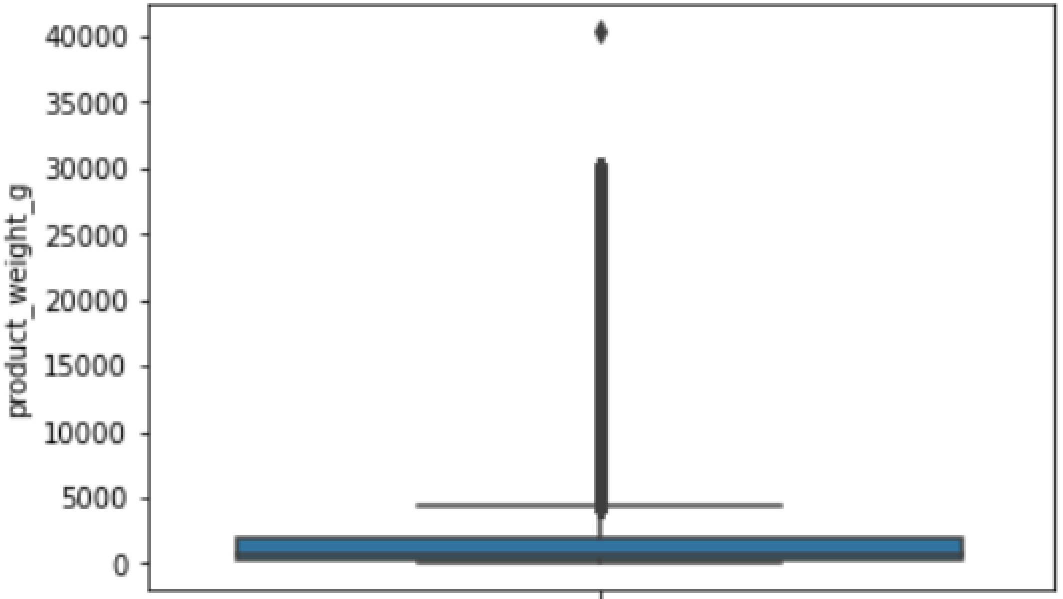
```
order.isna().sum().sort_values(ascending= False)

order_id                         0
customer_id                      0
order_status                     0
order_purchase_timestamp         0
order_approved_at                0
order_delivered_timestamp        0
order_estimated_delivery_date    0
dtype: int64
```

- Handled Missing values in sheet ‘products’ by filling the product_category_name by mode at zeroth position and got he insight as above:

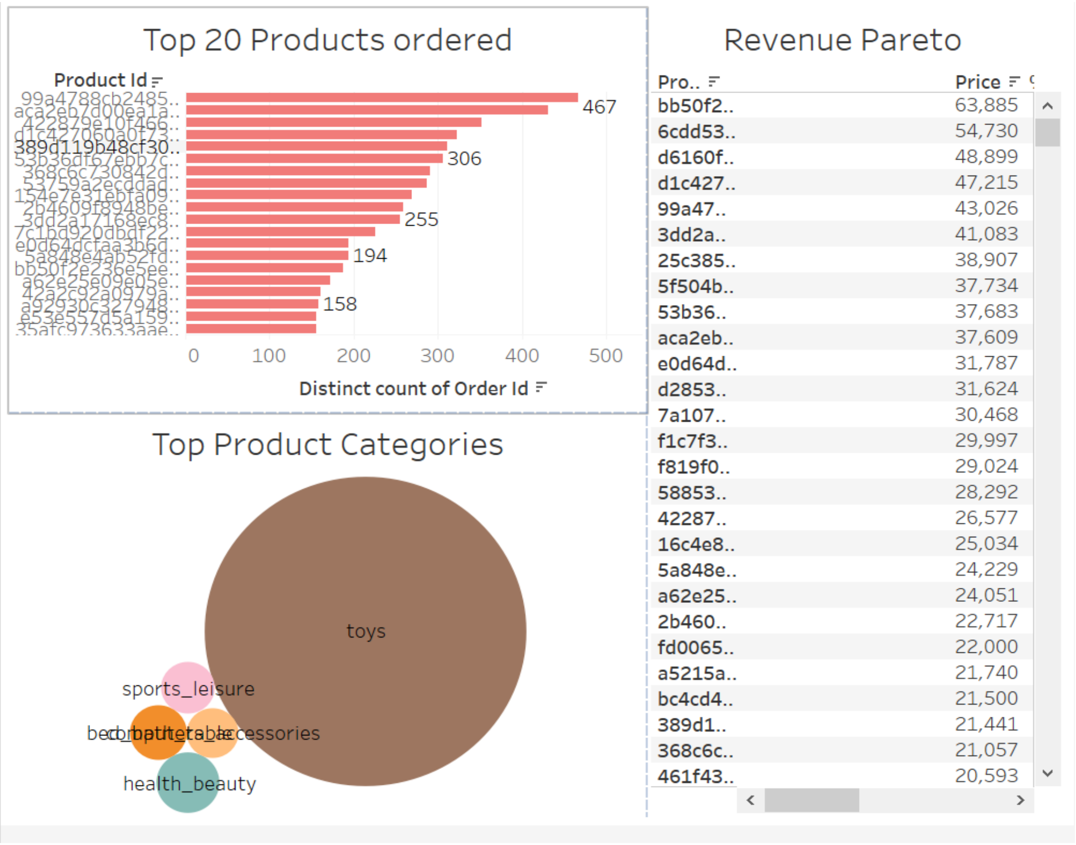
Outliers in product_weight_g category

- 1.5(IQR) rule = 1.5(Q3-Q1) = 1.5(1900-300) = 1.5*1600= 2400

- Lower outliers= Q1-2400 = 300-2400 = -2100
- Higher outliers= Q3+2400=1900+2400= 3300

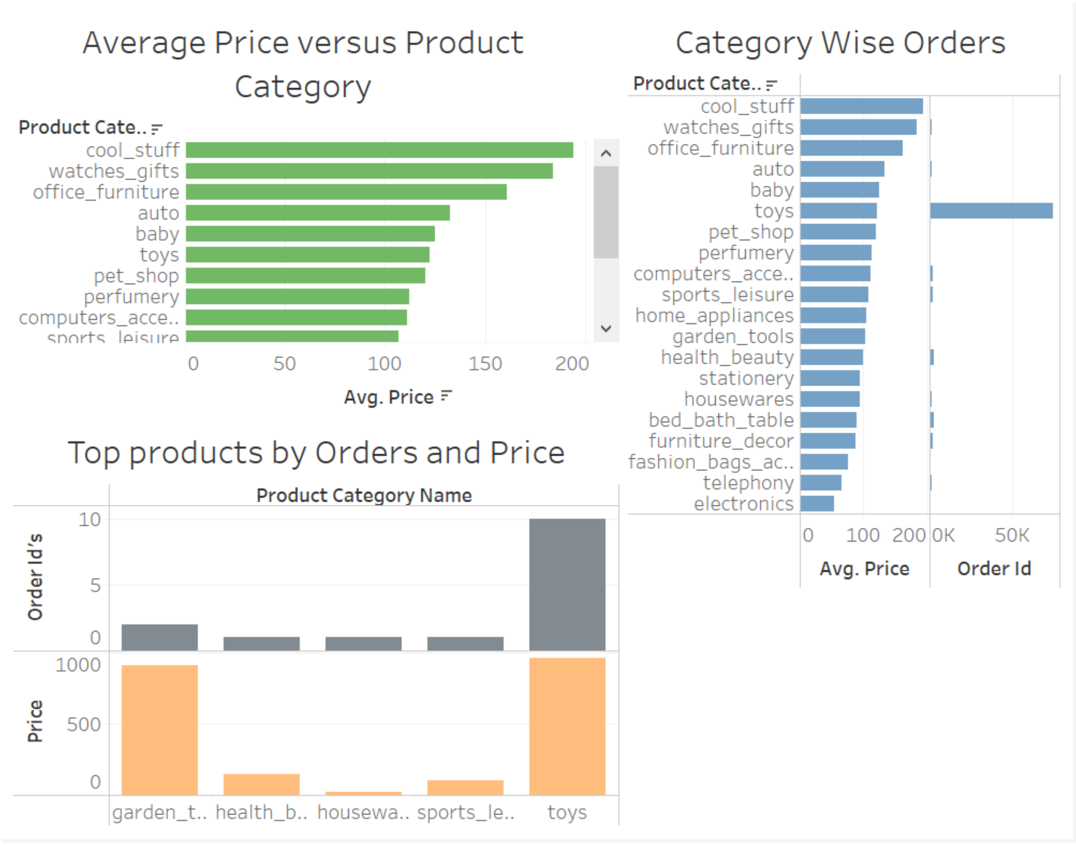


Analysis Using Tableau

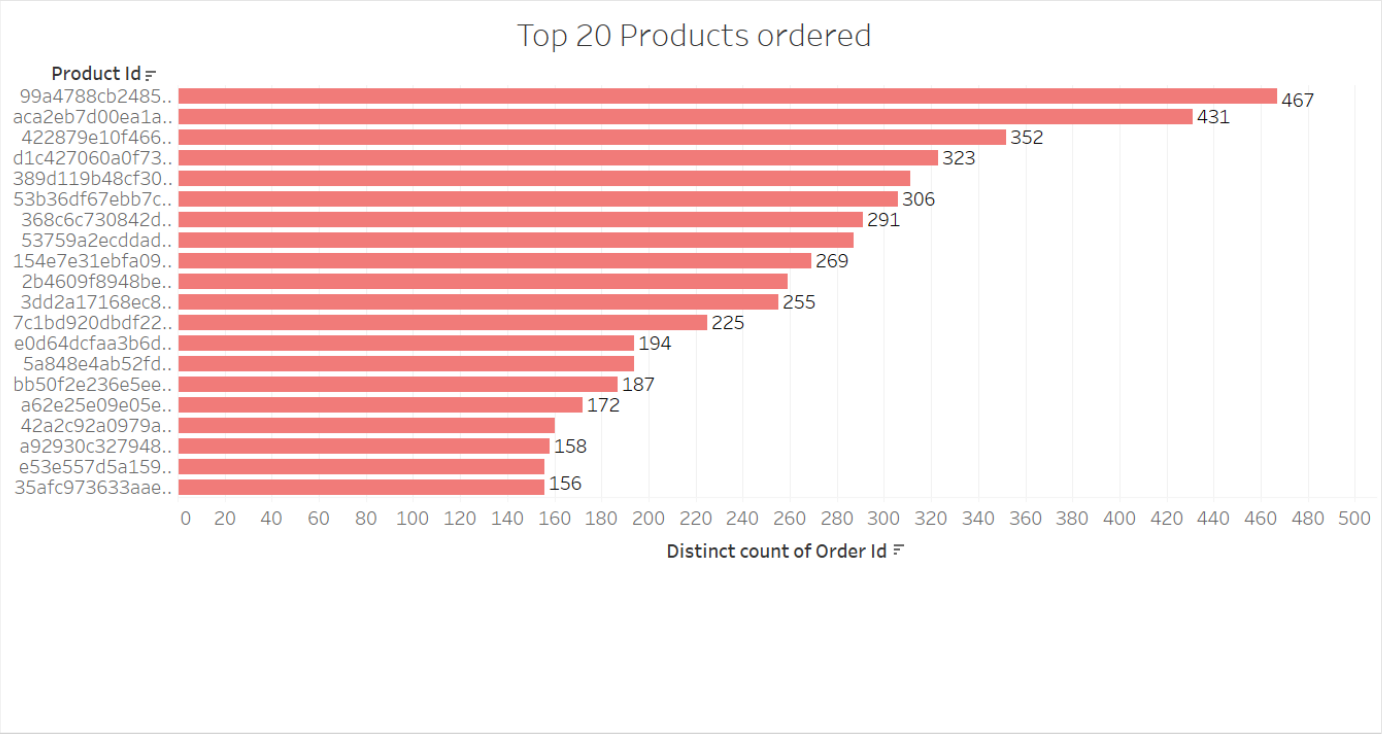


Revenue Pareto

Pro..	Price
bb50f2..	63,885
6cdd53..	54,730
d6160f..	48,899
d1c427..	47,215
99a47..	43,026
3dd2a..	41,083
25c385..	38,907
5f504b..	37,734
53b36..	37,683
aca2eb..	37,609
e0d64d..	31,787
d2853..	31,624
7a107..	30,468
f1c7f3..	29,997
f819f0..	29,024
58853..	28,292
42287..	26,577
16c4e8..	25,034
5a848e..	24,229
a62e25..	24,051
2b460..	22,717
fd0065..	22,000
a5215a..	21,740
bc4cd4..	21,500
389d1..	21,441
368c6c..	21,057
461f43..	20,593



Top Products Ordered

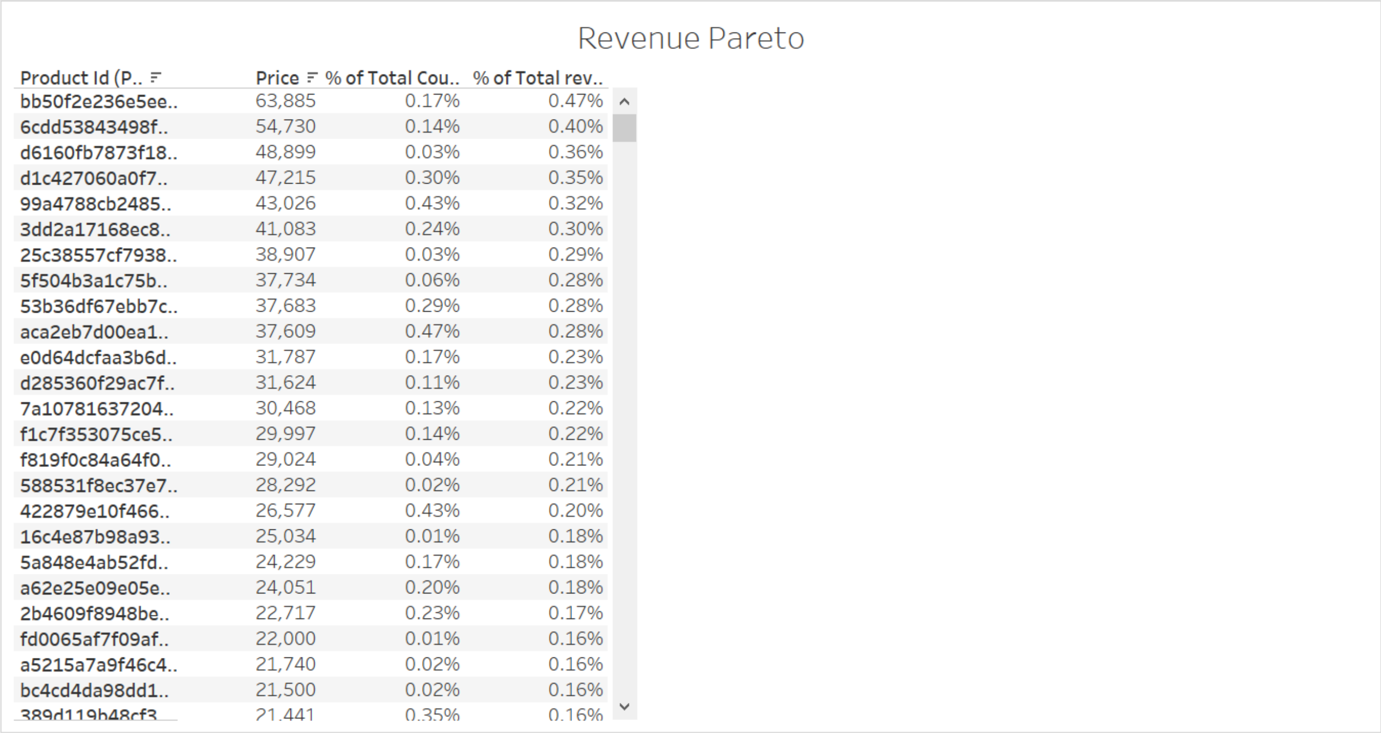


Top 20 Products ordered

Insights:

1. The encrypted product_ID corresponding to the mapped item are the top 20 items that are ordered.
2. If the company provided the decrypted names of all the items, then all of the names would have been arranged by top 20 standards.

Top Products and its revenue



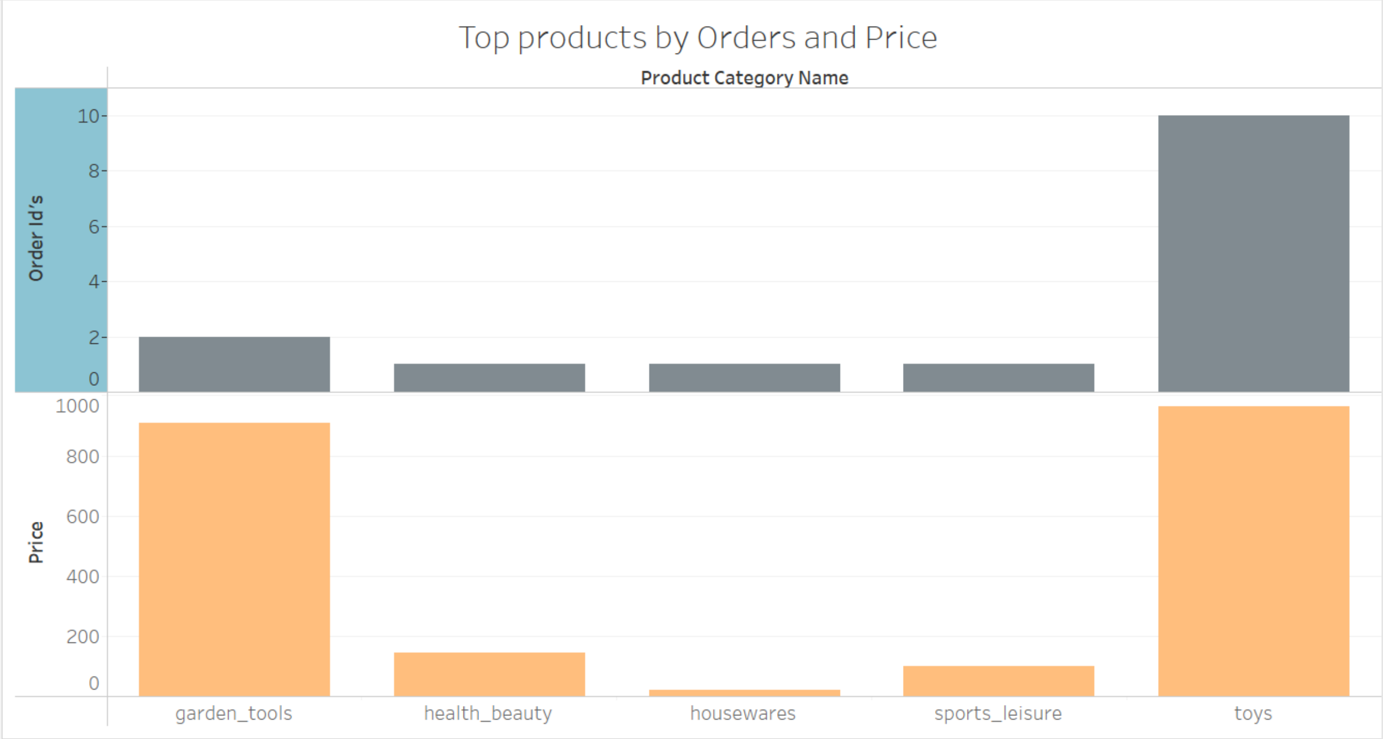
Revenue Pareto

Product Id (P..	Price	% of Total Cou..	% of Total rev..
bb50f2e236e5ee..	63,885	0.17%	0.47%
6cdd53843498f..	54,730	0.14%	0.40%
d6160fb7873f18..	48,899	0.03%	0.36%
d1c427060a0f7..	47,215	0.30%	0.35%
99a4788cb2485..	43,026	0.43%	0.32%
3dd2a17168ec8..	41,083	0.24%	0.30%
25c38557cf7938..	38,907	0.03%	0.29%
5f504b3a1c75b..	37,734	0.06%	0.28%
53b36df67ebb7c..	37,683	0.29%	0.28%
aca2eb7d00ea1..	37,609	0.47%	0.28%
e0d64dcfaa3b6d..	31,787	0.17%	0.23%
d285360f29ac7f..	31,624	0.11%	0.23%
7a10781637204..	30,468	0.13%	0.22%
f1c7f353075ce5..	29,997	0.14%	0.22%
f819f0c84a64f0..	29,024	0.04%	0.21%
588531f8ec37e7..	28,292	0.02%	0.21%
422879e10f466..	26,577	0.43%	0.20%
16c4e87b98a93..	25,034	0.01%	0.18%
5a848e4ab52fd..	24,229	0.17%	0.18%
a62e25e09e05e..	24,051	0.20%	0.18%
2b4609f8948be..	22,717	0.23%	0.17%
fd0065af7f09af..	22,000	0.01%	0.16%
a5215a7a9f46c4..	21,740	0.02%	0.16%
bc4cd4da98dd1..	21,500	0.02%	0.16%
389d119b48cf3..	21,441	0.35%	0.16%

Insights:

The encrypted product_ID corresponding to the mapped item are the top 20 items that are ordered by price, the total count of orders and lastly total revenue generated as sum of price.

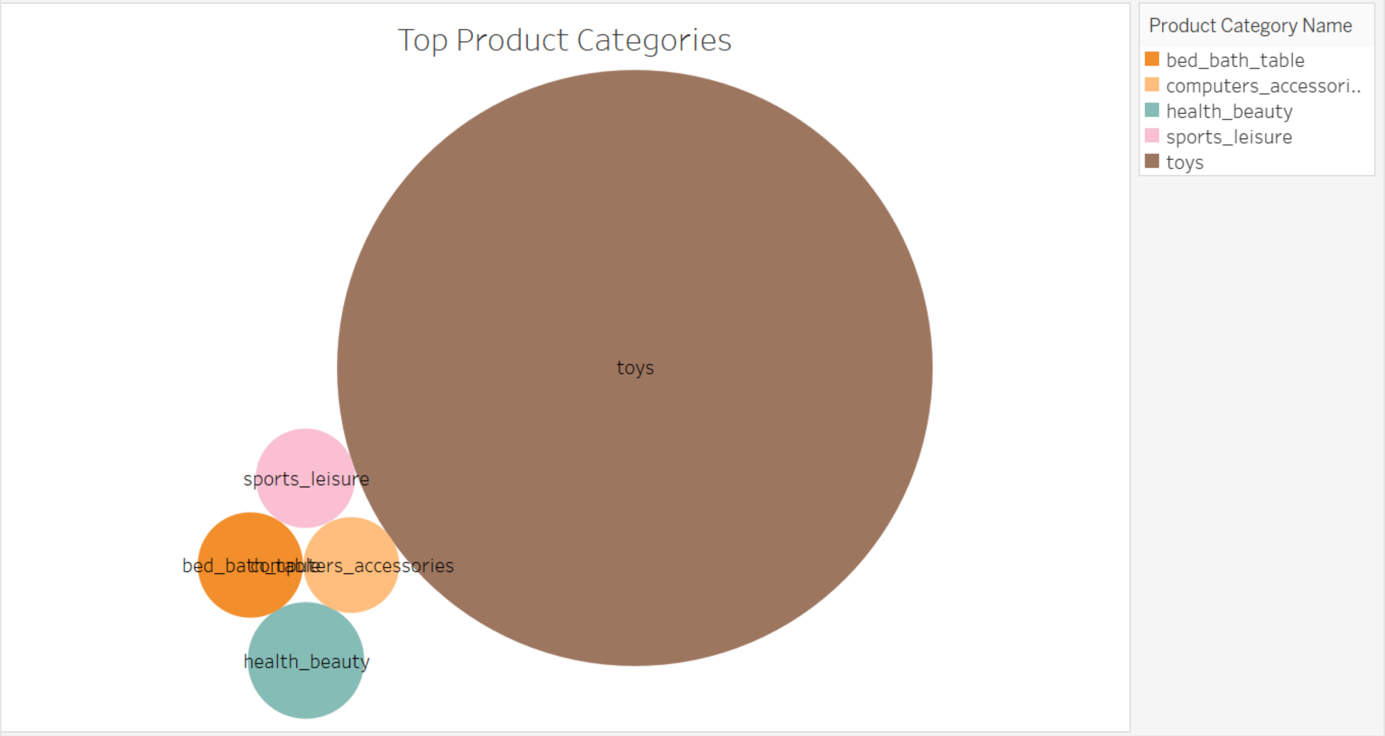
Top Products by Number of Orders and Price



Insights:

1. We can clearly see Toys have the most order_id's and Price
2. Whereas Garden_items can be considered as costly but not ordered so frequently

Top Product categories as per Count of order ID’s

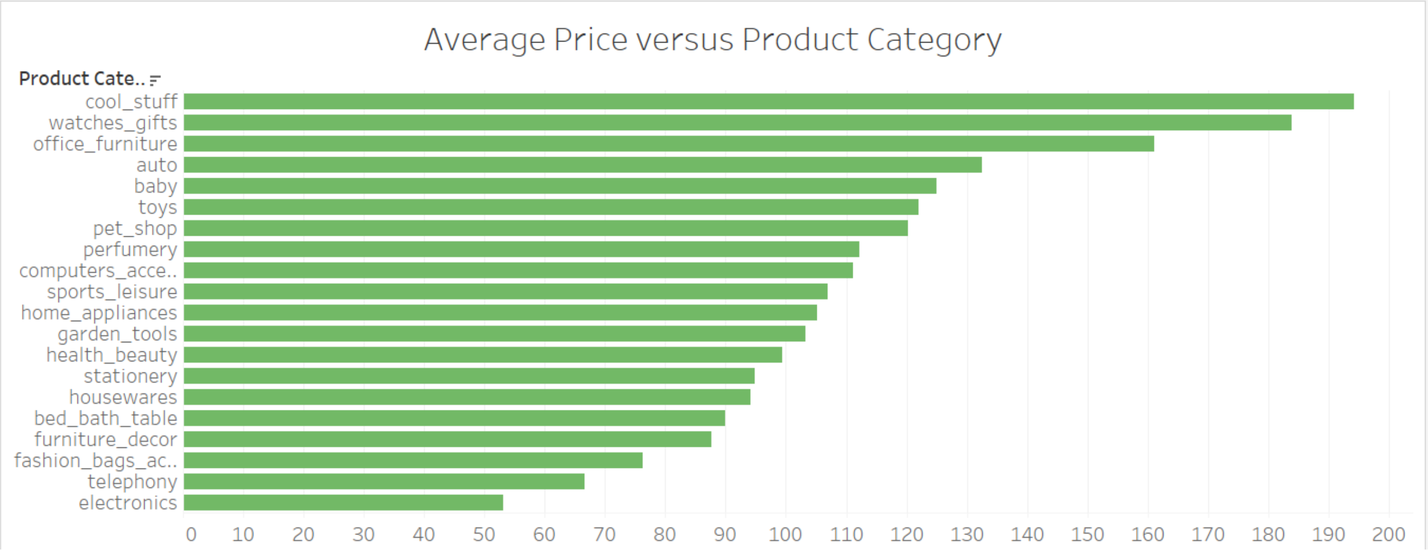


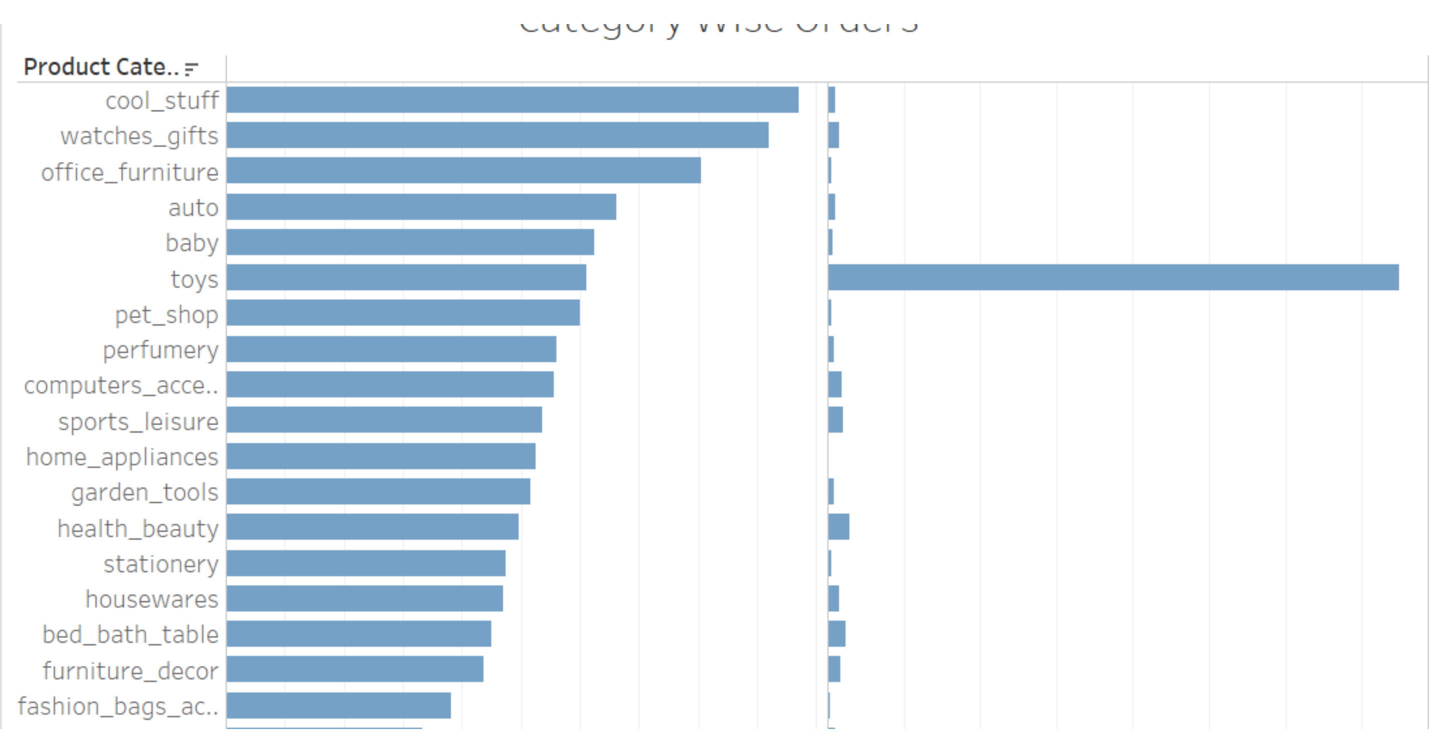
Insights:

1. Clearly we can see Toys have the most orders by order_category followed by health and beauty

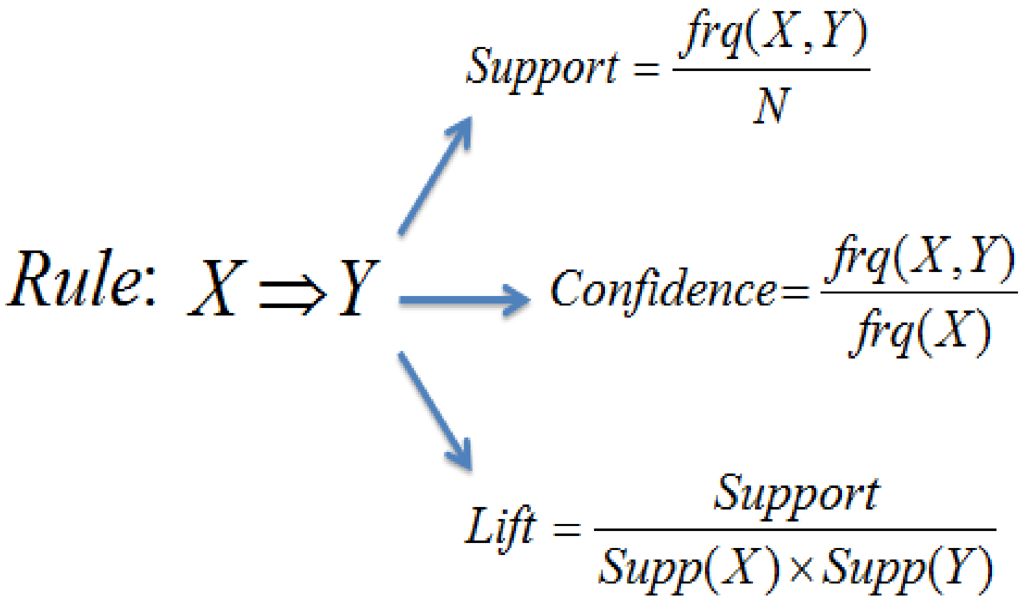
Top 20 Product Category versus average price

Clearly, we can see that product category as Cool_stuff has highest acceptability across all the other categories.





Market basket analysis in Python



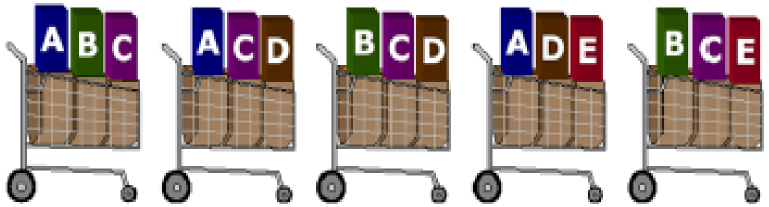
$$Rule:\ X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Example:

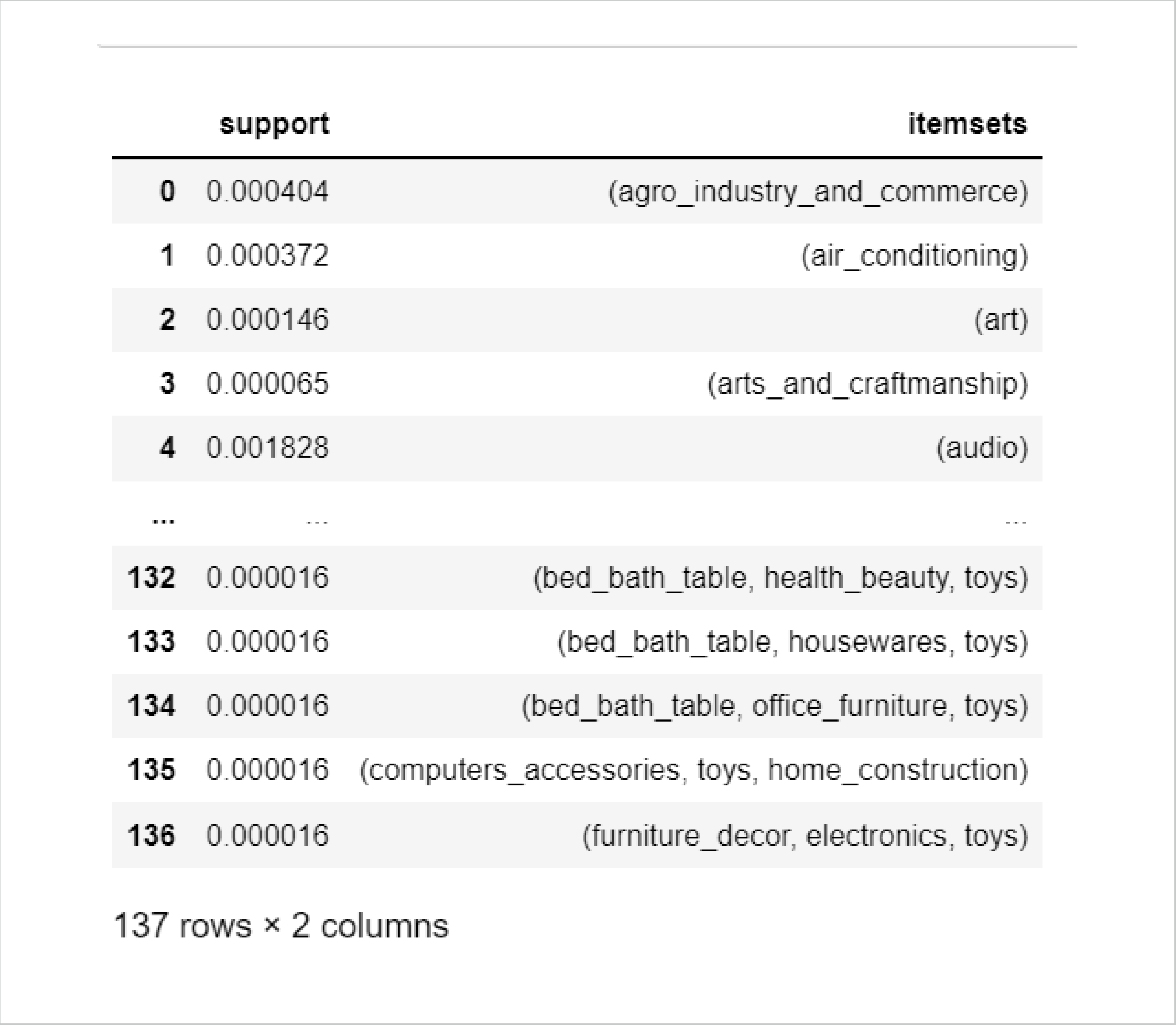


Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \,\&\, C \Rightarrow D$	1/5	1/3	5/9

Market basket analysis in Python

- Itemsets and its support

Support: > This measure gives an idea of how frequent ItemSet is in all the transactions. Example (agro_industry_and_commerce)is present in 0.0404% of the transactions



	support	itemsets
0	0.000404	(agro_industry_and_commerce)
1	0.000372	(air_conditioning)
2	0.000146	(art)
3	0.000065	(arts_and_craftmanship)
4	0.001828	(audio)
...	...	...
132	0.000016	(bed_bath_table, health_beauty, toys)
133	0.000016	(bed_bath_table, housewares, toys)
134	0.000016	(bed_bath_table, office_furniture, toys)
135	0.000016	(computers_accessories, toys, home_construction)
136	0.000016	(furniture_decor, electronics, toys)

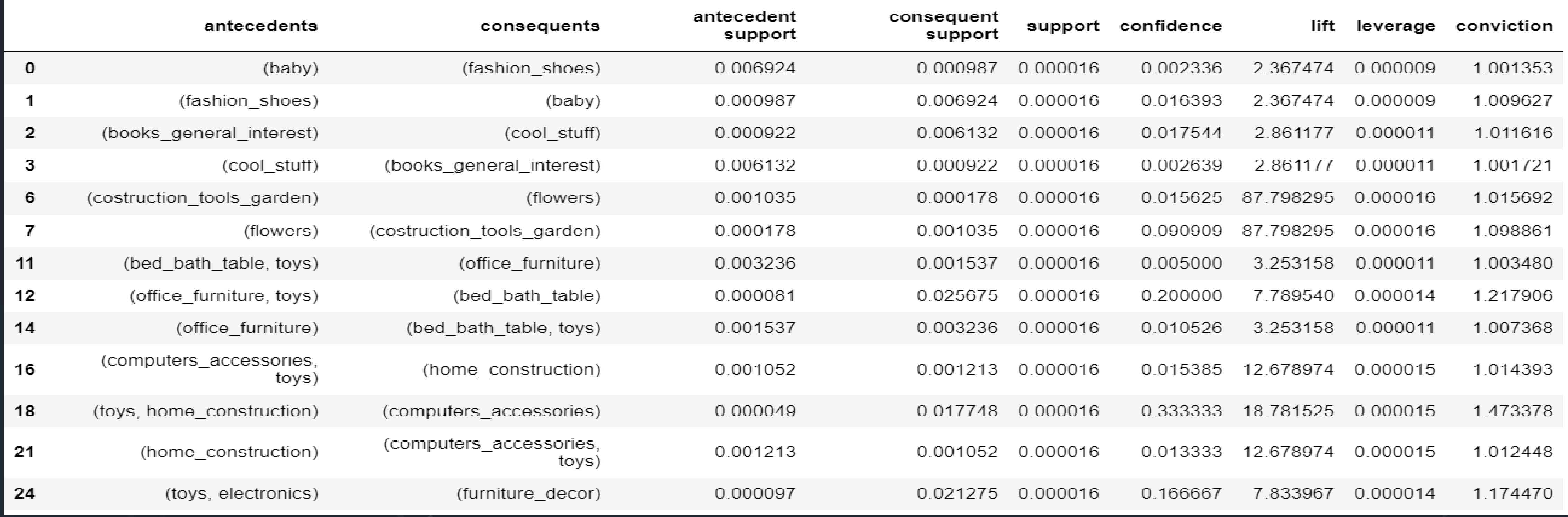
137 rows × 2 columns

Important terms:

Confidence: This measure defines the likeliness of occurrence of consequent on the cart given that the cart already has the antecedents.

Lift: This measure defines the likeliness of occurrence of consequent on the cart given that the cart already has the antecedent but controlling the popularity of consequent. For the items more than 1 is worth considering.

INSIGHTS



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(baby)	(fashion_shoes)	0.006924	0.000987	0.000016	0.002336	2.367474	0.000009	1.001353
1	(fashion_shoes)	(baby)	0.000987	0.006924	0.000016	0.016393	2.367474	0.000009	1.009627
2	(books_general_interest)	(cool_stuff)	0.000922	0.006132	0.000016	0.017544	2.861177	0.000011	1.011616
3	(cool_stuff)	(books_general_interest)	0.006132	0.000922	0.000016	0.002639	2.861177	0.000011	1.001721
6	(costruction_tools_garden)	(flowers)	0.001035	0.000178	0.000016	0.015625	87.798295	0.000016	1.015692
7	(flowers)	(costruction_tools_garden)	0.000178	0.001035	0.000016	0.090909	87.798295	0.000016	1.098861
11	(bed_bath_table, toys)	(office_furniture)	0.003236	0.001537	0.000016	0.005000	3.253158	0.000011	1.003480
12	(office_furniture, toys)	(bed_bath_table)	0.000081	0.025675	0.000016	0.200000	7.789540	0.000014	1.217906
14	(office_furniture)	(bed_bath_table, toys)	0.001537	0.003236	0.000016	0.010526	3.253158	0.000011	1.007368
16	(computers_accessories, toys)	(home_construction)	0.001052	0.001213	0.000016	0.015385	12.678974	0.000015	1.014393
18	(toys, home_construction)	(computers_accessories)	0.000049	0.017748	0.000016	0.333333	18.781525	0.000015	1.473378
21	(home_construction)	(computers_accessories, toys)	0.001213	0.001052	0.000016	0.013333	12.678974	0.000015	1.012448
24	(toys, electronics)	(furniture_decor)	0.000097	0.021275	0.000016	0.166667	7.833967	0.000014	1.174470

- **Support**: This measure gives an idea of how frequent `ItemSet` is in all the transactions. Like (fashion_shoes),(baby)} is present in 0.0015% of the transactions.
- **Confidence**: This measure defines the likeliness of occurrence of consequent on the cart given that the cart already has the antecedents. In simple words there is an 1.63% chance of finding {fashion_shoes} , if the cart contains {baby}.

  Top 2 Categories will be: (electronics, toys)-> (electronics, toys) with 16.66% confidence and

  (home_construction, toys) -> (computers_accessories) with 25.00% confidence.
- **Lift:** For the items more than 1 is worth considering which means this measure defines the likeliness of occurrence of consequent on the cart given that the cart already has the antecedent but controlling the popularity of consequent. So, lift of (fashion_shoes) w.r.t (baby) is 2.24. Which is quite good. Any lift value > 1 implies that the Association rule is worth considering.

Thank you