BRAZILIAN ECOMMERCE

Lab Group: FEP2
Team Number: 03
Alexander Ng
Chan Eu Ching
Mitra Ren Sachi

PROBLEM

Given a product category, are we able to predict other categories that are likely purchased together?

TABLE OF CONTENTS

“Malloc allocates from the HEAP.”

—PIKA PIKA

01

DATA CLEANING

Steps for Data Cleaning

1. Importing .csv files
2. Extracting columns out from individual .csv files into dataframes
3. Merge all the dataframes together
4. Organize columns
    a. Dropping columns (customer_id, product_category_name)
    b. Renaming product_category_name_english to product_category_name
    c. Reordering columns
5. Sort according to highest customer_unique_id occurrence (most active customer)

Things to note:

→ Checking the count of every dataframe collected, to ensure that there are no NULL values

→ When merging dataframes, ensure that count for unique customer_unique_id remains constant

DATASETS USED

OLIST_ORDERS_DATASET.CSV

order_id
customer_id
order_purchase_timestamp

OLIST_ORDER_ITEMS_DATASET.CSV

order_id
order_item_id
product_id
seller_id

OLIST_CUSTOMER_DATASET.CSV

customer_id
customer_unique_id
customer_state

OLIST_PRODUCTS_DATASET.CSV

product_id
product_category_name

PRODUCT_CATEGORY_TRANSLATION.CSV

product_category_name
product_category_name_english

OLIST_SELLERS_DATSET.CSV

seller_id
seller_state

MAIN DATAFRAME

	customer_unique_id	customer_state	order_id	order_item_id	order_purchase_timestamp
				count	
56388	8d50f5eadf50201ccdcedfb9e2ac8455	SP	112eb6f37f1b9dabbced368fbbc6c9ef	1	2018-07-23 21:53:02
56389	8d50f5eadf50201ccdcedfb9e2ac8455	SP	23427a6bd9f8fd1b51f1b1e5cc186ab8	1	2018-05-21 22:44:31
56390	8d50f5eadf50201ccdcedfb9e2ac8455	SP	369634708db140c5d2c4e365882c443a	1	2017-06-18 22:56:48
56391	8d50f5eadf50201ccdcedfb9e2ac8455	SP	4f62d593acae92cea3c5662c76122478	1	2017-07-18 23:10:58
56392	8d50f5eadf50201ccdcedfb9e2ac8455	SP	519203404f6116d406a970763ee75799	1	2017-08-05 08:59:43
...	...	...	...	...	...
27088	43c272f80acfc8b161137776e983cae3	RJ	59a5648d91a60574c541152e2fe1c66c	1	2017-09-12 13:52:02
17831	2c98316ef32eba5280f5d4cf2a505d0f	RJ	158a05ea7d6a7268559ca2f97476ee0a	1	2017-09-14 10:30:22
23987	3be9c553e23e8c83ced423a32f5ea1f0	SP	3f7ac2772804d1316e995c3621024c7f	1	2018-03-11 11:38:50
67824	a966fbec25ff8c8540fa2e3c9076ac88	SP	7656e54ae9322214b7459746ca6076a8	1	2018-05-03 22:15:36
40724	6597bb764c5cb4bee6ee7509c1c61b64	RJ	4ed72b48d55c5cb2ad3a63f2b96dca12	1	2017-12-01 18:02:11

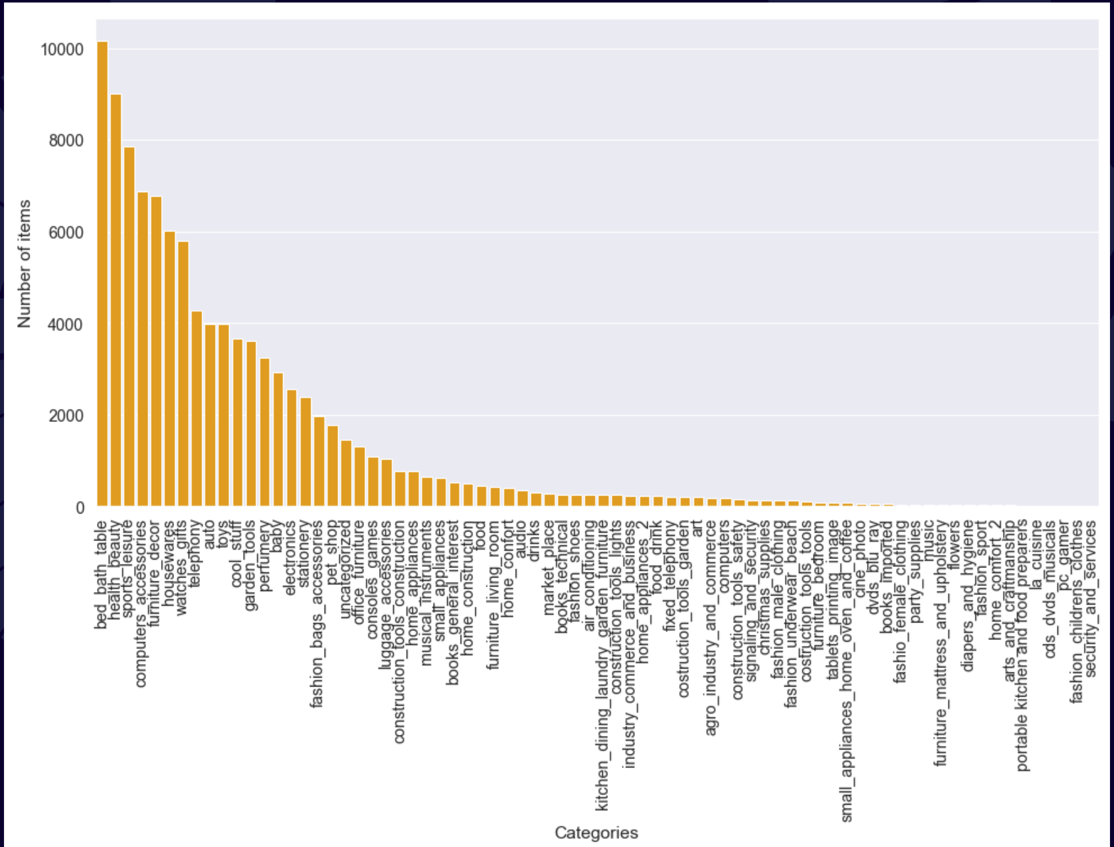
102403 rows × 9 columns

product_id	product_category_name	seller_id	seller_state
41f6cb7c3b1200749326e50106f32d58	sports_leisure	db4350fd57ae30082dec7acbaacc17f9	SP
5cb96c51c55f57503465e4d2558dc053	sports_leisure	db4350fd57ae30082dec7acbaacc17f9	SP
d83509907a19c72e1e4cdde78b8177ec	sports_leisure	94e93ce877be27a515118dbfd2c2be41	SP
94cc774056d3f2b0dc693486a589025e	fashion_bags_accessories	1da3aeb70d7989d1e6d9b0e887f97c23	SP
5fb61f482620cb672f5e586bb132eae9	uncategorized	94e93ce877be27a515118dbfd2c2be41	SP
...	...	...	...
d52d7fb0d4ea10cd52baa3255c5c0a34	sports_leisure	e1b12447a7563944843191754aeb5562	SP
34dabb8af33b3756cf72df05fb327011	tablets_printing_image	0db783cfcd3b73998abc6e10e59a102f	SP
89321f94e35fc6d7903d36f74e351d40	food	16090f2ca825584b5a147ab24aa30c86	SP
61d01171a3784bab50137290350e5332	housewares	4992e76a42cb3aad7a047e0d3d7e729	SP
252641aa4855aef622089db60c4ad90a	cool_stuff	06e5eefc71ec47ae763c5c6f8db7064f	RS

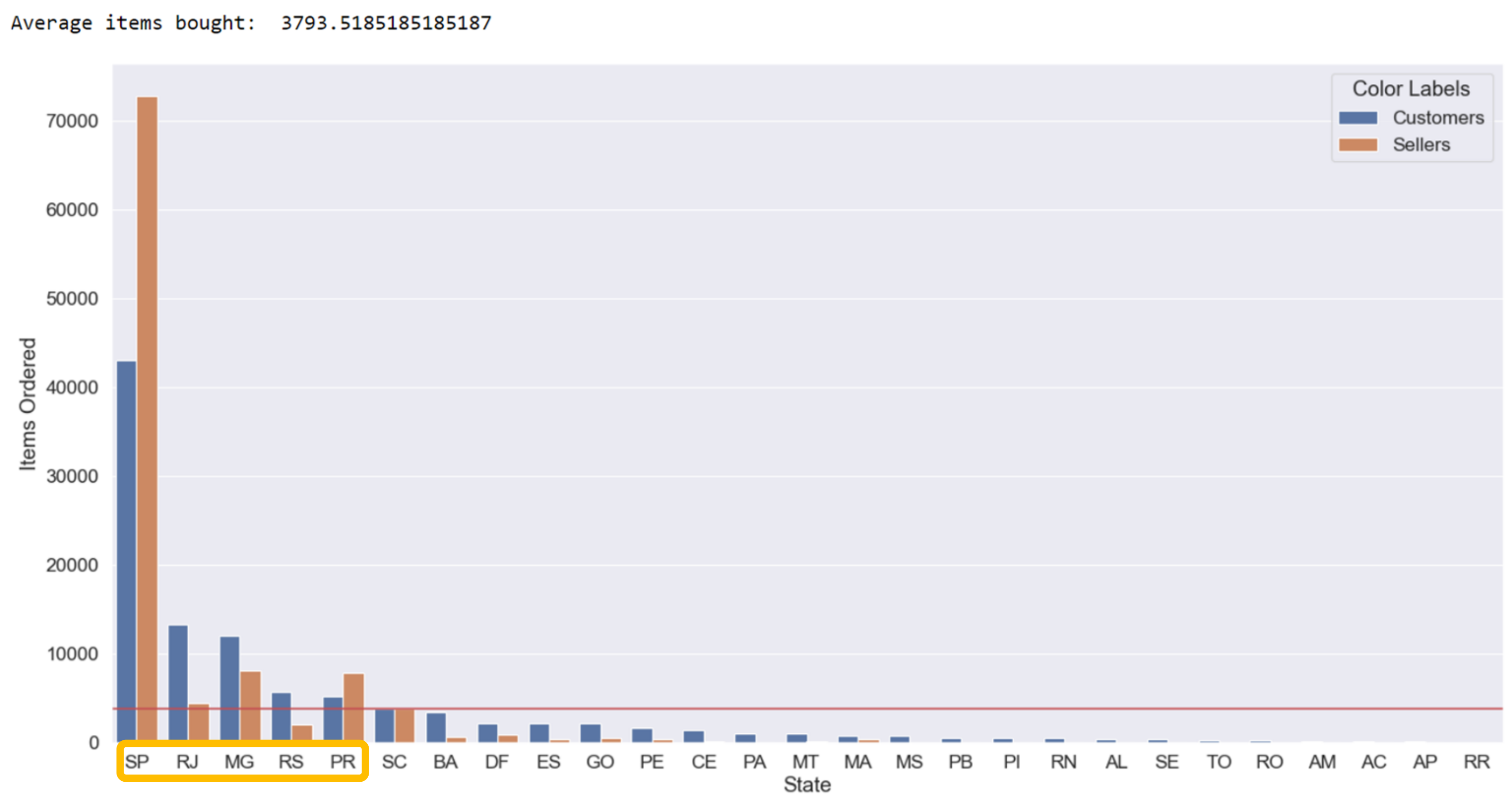
02

EXPLORATORY DATA ANALYSIS

DISTRIBUTION OF ITEM CATEGORIES



ITEMS BOUGHT PER STATE



DISTRIBUTION OF ITEMS BOUGHT IN CATEGORIES PER STATE

SP: São Paulo

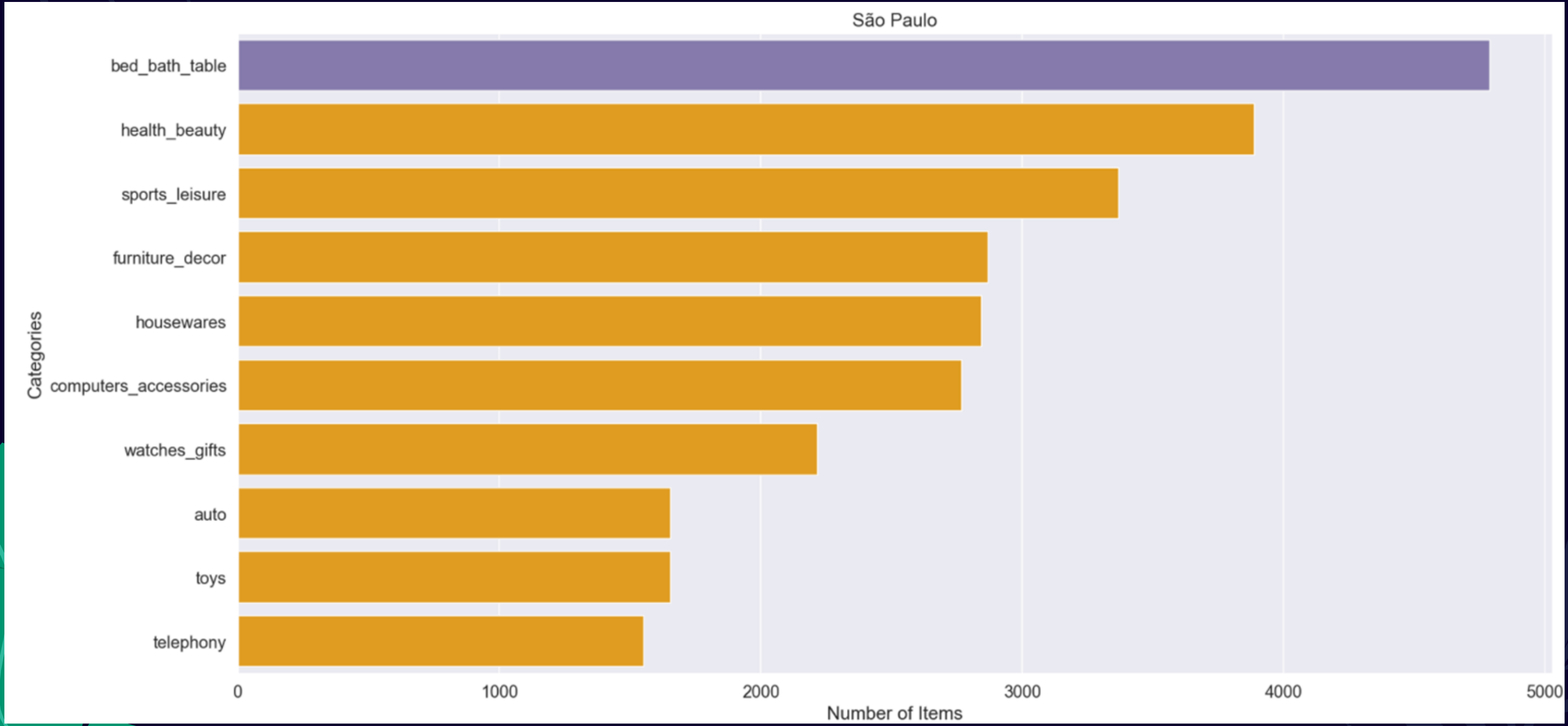
MG: Minas Gerais

RJ: Rio de Janeiro

PR: Paraná

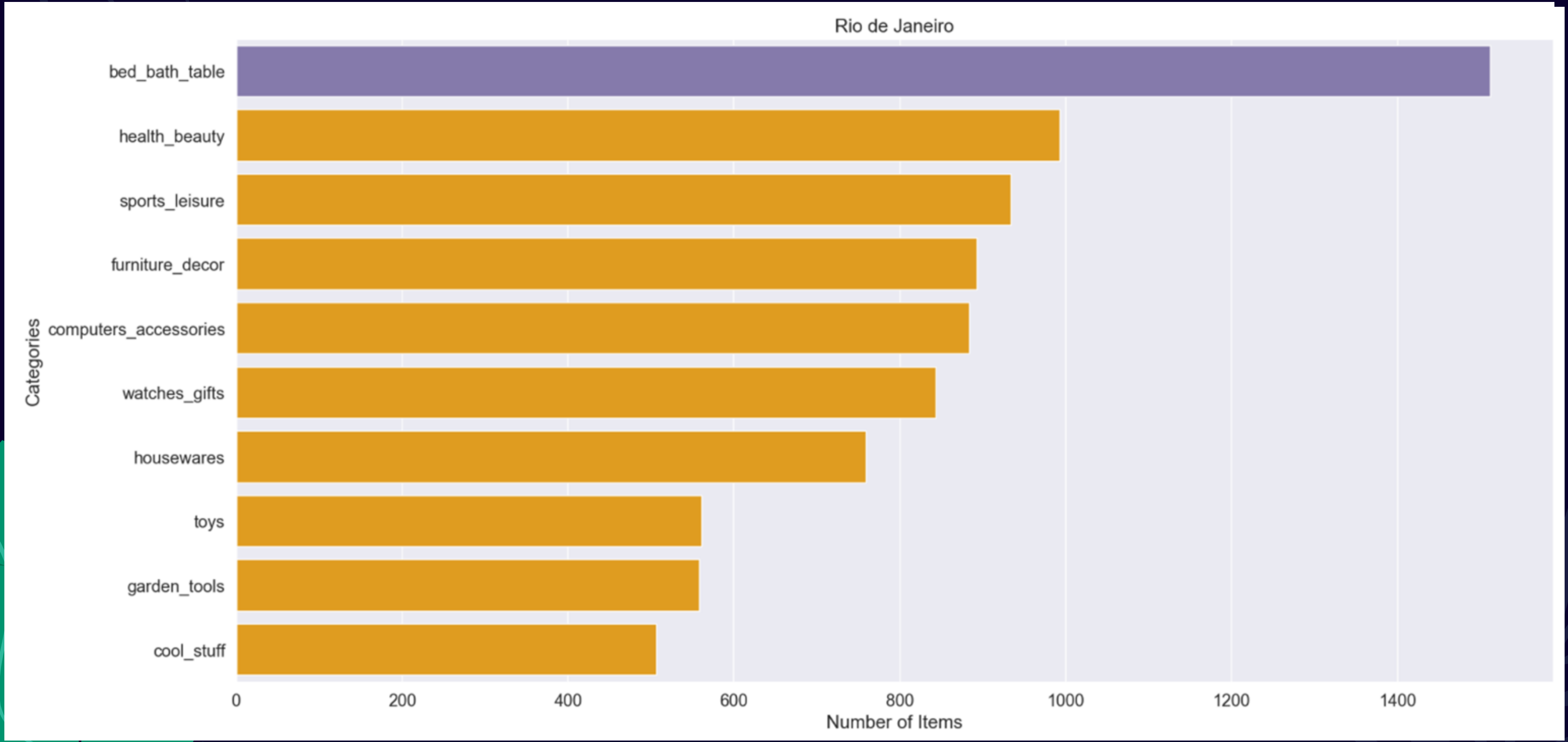
RS: Rio Grande do Sul

SP: SÃO PAULO

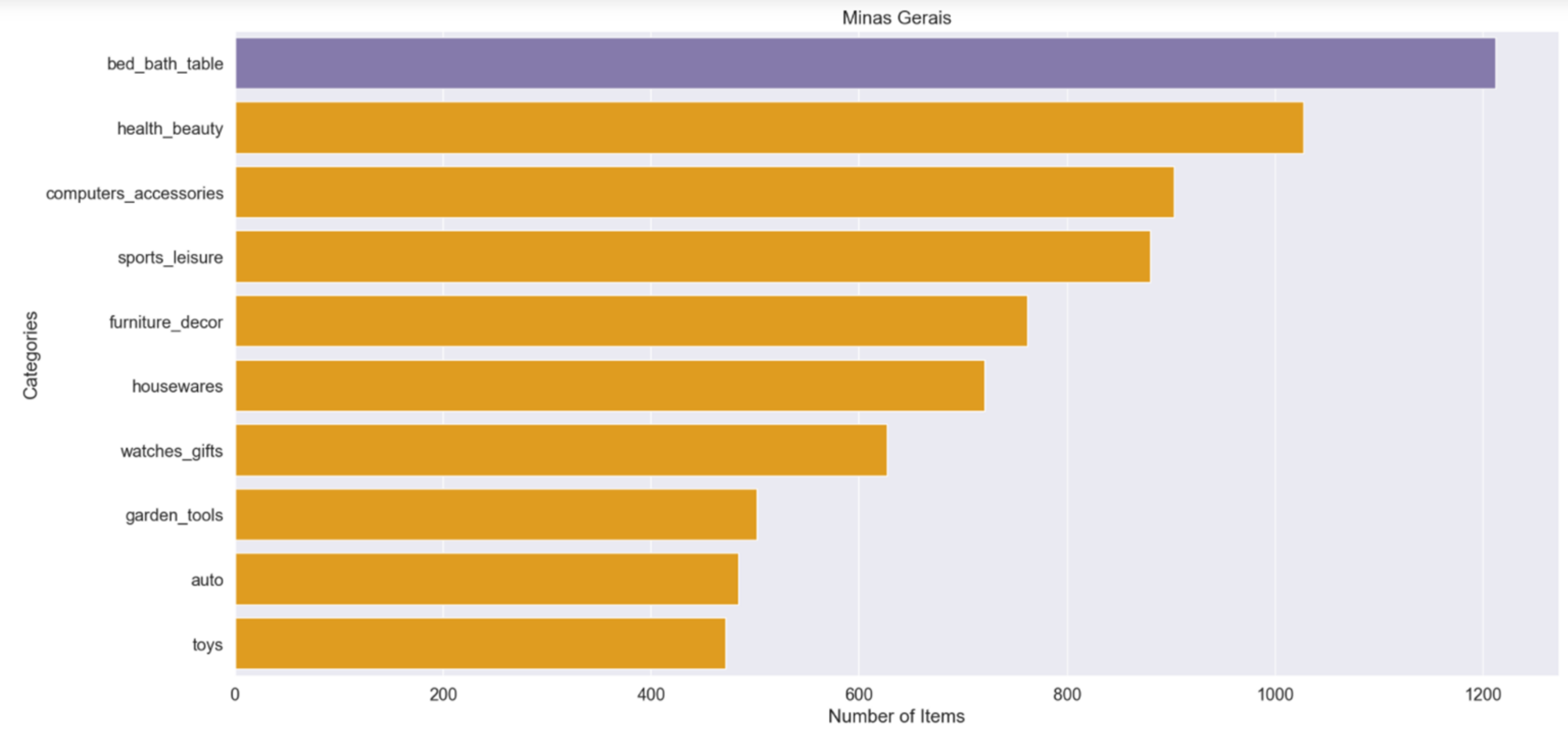


São Paulo
bed_bath_table
health_beauty
sports_leisure
furniture_decor
housewares
computers_accessories
watches_gifts
auto
toys
telephony
Categories
0
1000
2000
3000
4000
5000
Number of Items

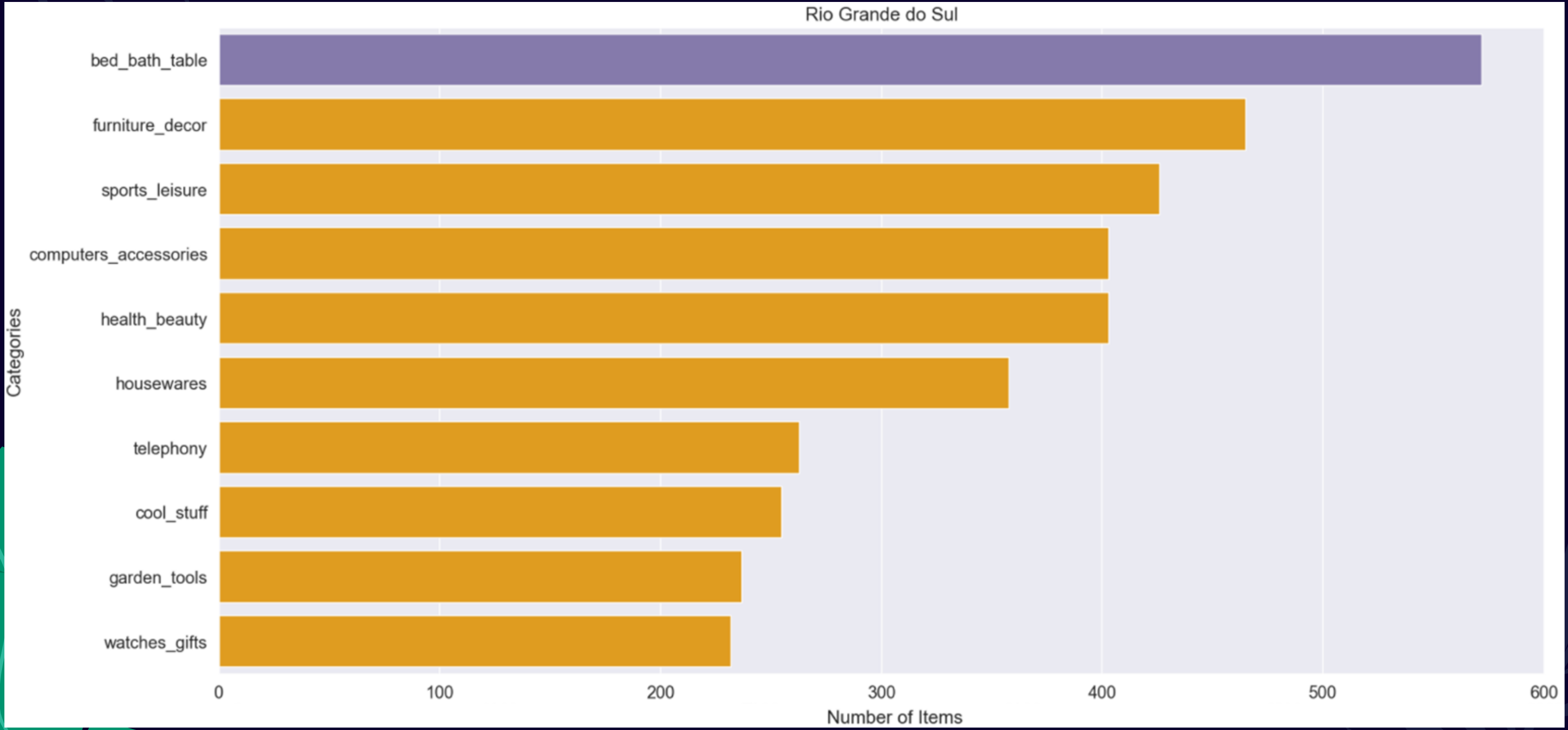
RJ: Rio de Janeiro



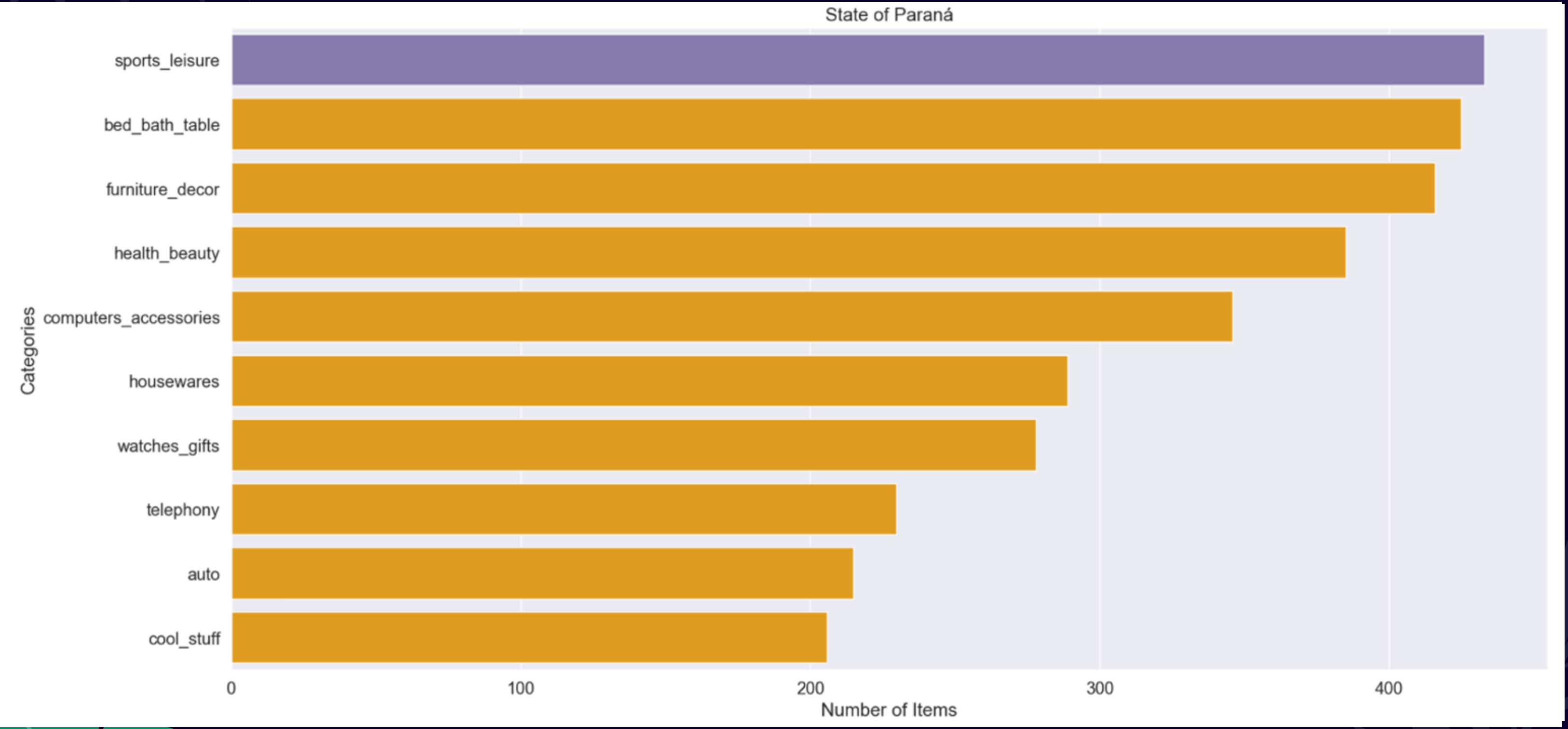
MG: Minas Gerais



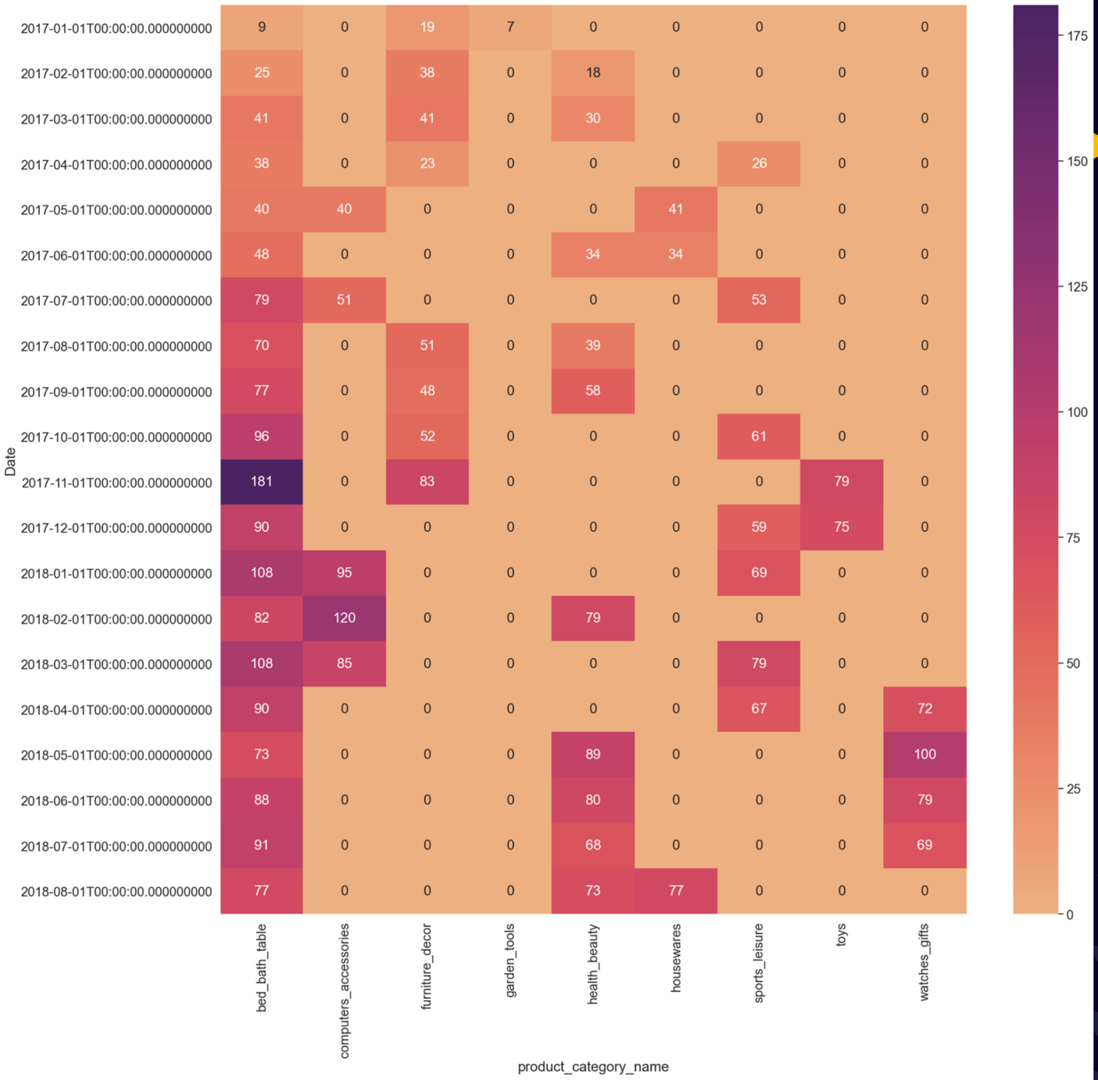
RS: RIO GRANDE DO SUL



PR: PARANÁ


State of Paraná
sports_leisure
bed_bath_table
furniture_decor
health_beauty
computers_accessories
housewares
watches_gifts
telephony
auto
cool_stuff
Categories
0
100
200
300
400
Number of Items

TOP 3 CATEGORIES OVER 2 YEARS



INSIGHTS (UWU)

Linking back to problem

Problem: Given a product category, are we able to predict other categories that are likely purchased together?

In EDA, firstly, we explore the top 10 categories of each of the five states to look for any trends that could be because of geolocation, general industry, and behavior of the locals

Secondly, we explore the top 3 categories for one of the most populated state, Rio de Janeiro, from 2017 to 2018 check for any time based abnormalities which could be because of festivals, or events that took place at that time.

03

MACHINE LEARNING

BACKGROUND

ASSOCIATION RULE LEARNING

Draws associations with items in a dataset and is an important concept of machine learning being used in market basket analysis

APPLICATION

- In the Olist E-commerce site, products are organised in terms of categories
- Investing time and resources on deliberate product placements reduces a customer's shopping time and reminds the customer of what relevant items they might be interested in buying
- Helps Olist cross-sell across product categories in the process.

Disclaimer!

Due to the ambiguous nature of the datasets available with regard to product names, we will group the products according categories as these labels are readily available

ARL MODEL

AIM

To show the relations between product categories in the Olist dataset

ALGORITHM

Uses a bottom-up approach where frequent items are extended one item at a time and groups of candidates are tested against the available dataset

INPUT

Binary table of order_id against product_category_name. Each row in the table represents a “market basket” belonging to a unique order transaction

OUTPUT

A table of rules that show which product categories have high associations with another

Evaluation metrics

Support

Proportion of a category

Confidence

Measures how often a consequent category appears in transactions that contain a given an antecedent

Lift

How likely a consequent category is bought together with a given antecedent

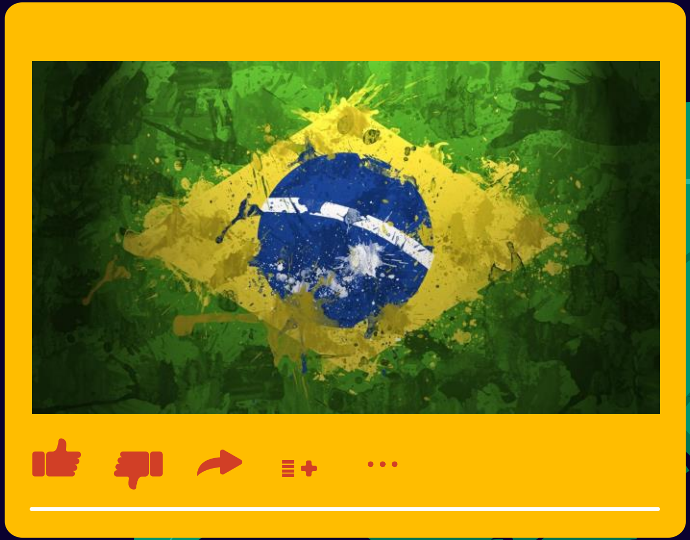
$$\text{Support (Item I)} = \frac{(Total\ Number\ of\ transactions\ with\ Item\ I)}{(Total\ number\ of\ transactions)}$$

$$\text{Confidence (Item } I_1 \rightarrow \text{Item } I_2\text{)} = \frac{No.\ of\ transactions\ with\ I_1\ and\ I_2}{No.\ of\ transactions\ with\ I_1}$$

$$\text{Lift (Item } I_1 \rightarrow \text{Item } I_2\text{)} = \frac{Confidence(Item\ I_1 \rightarrow I_2)}{Supprt(Item\ I_2)}$$

Association Rules for the Entire Brazil

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
2	((, bed_bath_table))	((, home_confort))	0.095443	0.004024	0.000436	0.004566	1.134835
3	((, home_confort))	((, bed_bath_table))	0.004024	0.095443	0.000436	0.108312	1.134835
12	((, furniture_decor))	((, construction_tools_lights))	0.065362	0.002473	0.000111	0.001706	0.689728
13	((, construction_tools_lights))	((, furniture_decor))	0.002473	0.065362	0.000111	0.045082	0.689728
10	((, home_construction))	((, furniture_decor))	0.004966	0.065362	0.000132	0.026531	0.405903
11	((, furniture_decor))	((, home_construction))	0.065362	0.004966	0.000132	0.002016	0.405903
4	((, baby))	((, cool_stuff))	0.029240	0.036811	0.000203	0.006932	0.188324
5	((, cool_stuff))	((, baby))	0.036811	0.029240	0.000203	0.005507	0.188324
6	((, toys))	((, baby))	0.039385	0.029240	0.000193	0.004889	0.167214
7	((, baby))	((, toys))	0.029240	0.039385	0.000193	0.006586	0.167214
9	((, uncategorized))	((, housewares))	0.014706	0.059636	0.000142	0.009649	0.161791
8	((, housewares))	((, uncategorized))	0.059636	0.014706	0.000142	0.002379	0.161791
0	((, furniture_decor))	((, bed_bath_table))	0.065362	0.095443	0.000709	0.010854	0.113726
1	((, bed_bath_table))	((, furniture_decor))	0.095443	0.065362	0.000709	0.007433	0.113726



ASSOCIATION RULES FOR SAO PAULO

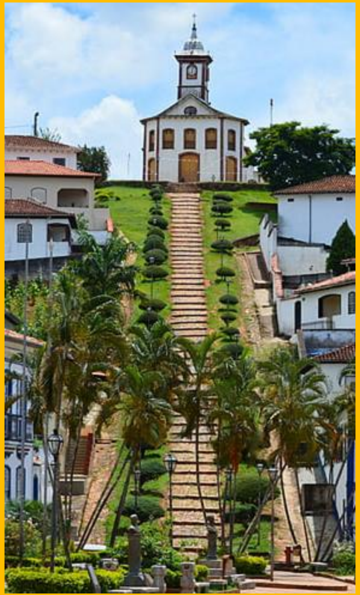
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	((, bed_bath_table))	((, home_confort))	0.106731	0.004447	0.000411	0.003850	0.865645
1	((, home_confort))	((, bed_bath_table))	0.004447	0.106731	0.000411	0.092391	0.865645
6	((, home_construction))	((, furniture_decor))	0.004931	0.065837	0.000145	0.029412	0.446737
7	((, furniture_decor))	((, home_construction))	0.065837	0.004931	0.000145	0.002203	0.446737
3	((, baby))	((, toys))	0.028616	0.038961	0.000242	0.008446	0.216781
2	((, toys))	((, baby))	0.038961	0.028616	0.000242	0.006203	0.216781
4	((, housewares))	((, uncategorized))	0.067215	0.014236	0.000193	0.002877	0.202075
5	((, uncategorized))	((, housewares))	0.014236	0.067215	0.000193	0.013582	0.202075
8	((, baby))	((, cool_stuff))	0.028616	0.031758	0.000145	0.005068	0.159567
9	((, cool_stuff))	((, baby))	0.031758	0.028616	0.000145	0.004566	0.159567

Association Rules for Rio de Janeiro

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
2	((, bed_bath_table))	((, home_confort))	0.109152	0.004388	0.000627	0.005743	1.308789
3	((, home_confort))	((, bed_bath_table))	0.004388	0.109152	0.000627	0.142857	1.308789
15	((, furniture_living_room))	((, office_furniture))	0.007522	0.017004	0.000157	0.020833	1.225230
14	((, office_furniture))	((, furniture_living_room))	0.017004	0.007522	0.000157	0.009217	1.225230
19	((, watches_gifts))	((, audio))	0.064018	0.004231	0.000157	0.002448	0.578539
18	((, audio))	((, watches_gifts))	0.004231	0.064018	0.000157	0.037037	0.578539
6	((, baby))	((, cool_stuff))	0.027033	0.039414	0.000392	0.014493	0.367707
7	((, cool_stuff))	((, baby))	0.039414	0.027033	0.000392	0.009940	0.367707
17	((, furniture_living_room))	((, furniture_decor))	0.007522	0.067231	0.000157	0.020833	0.309878
16	((, furniture_decor))	((, furniture_living_room))	0.067231	0.007522	0.000157	0.002331	0.309878
11	((, housewares))	((, baby))	0.058455	0.027033	0.000235	0.004021	0.148759
10	((, baby))	((, housewares))	0.027033	0.058455	0.000235	0.008696	0.148759
1	((, bed_bath_table))	((, furniture_decor))	0.109152	0.067231	0.001019	0.009332	0.138811
0	((, furniture_decor))	((, bed_bath_table))	0.067231	0.109152	0.001019	0.015152	0.138811
12	((, bed_bath_table))	((, uncategorized))	0.109152	0.015593	0.000235	0.002154	0.138113
13	((, uncategorized))	((, bed_bath_table))	0.015593	0.109152	0.000235	0.015075	0.138113
5	((, garden_tools))	((, furniture_decor))	0.043097	0.067231	0.000392	0.009091	0.135219
4	((, furniture_decor))	((, garden_tools))	0.067231	0.043097	0.000392	0.005828	0.135219
9	((, computers_accessories))	((, garden_tools))	0.067309	0.043097	0.000313	0.004657	0.108050
8	((, garden_tools))	((, computers_accessories))	0.043097	0.067309	0.000313	0.007273	0.108050

Association Rules for Minas Gerais

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
2	((, bed_bath_table))	((, home_confort))	0.097800	0.004591	0.000433	0.004429	0.964621
3	((, home_confort))	((, bed_bath_table))	0.004591	0.097800	0.000433	0.094340	0.964621
12	((, furniture_decor))	((, construction_tools_lights))	0.062976	0.003119	0.000173	0.002751	0.882164
13	((, construction_tools_lights))	((, furniture_decor))	0.003119	0.062976	0.000173	0.055556	0.882164
14	((, furniture_decor))	((, home_construction))	0.062976	0.005024	0.000173	0.002751	0.547550
15	((, home_construction))	((, furniture_decor))	0.005024	0.062976	0.000173	0.034483	0.547550
17	((, luggage_accessories))	((, stationery))	0.015246	0.021050	0.000173	0.011364	0.539843
16	((, stationery))	((, luggage_accessories))	0.021050	0.015246	0.000173	0.008230	0.539843
4	((, housewares))	((, uncategorized))	0.060811	0.014900	0.000347	0.005698	0.382429
5	((, uncategorized))	((, housewares))	0.014900	0.060811	0.000347	0.023256	0.382429
6	((, baby))	((, cool_stuff))	0.029886	0.037595	0.000347	0.011594	0.308395
7	((, cool_stuff))	((, baby))	0.037595	0.029886	0.000347	0.009217	0.308395
10	((, toys))	((, baby))	0.039674	0.029886	0.000260	0.006550	0.219176
11	((, baby))	((, toys))	0.029886	0.039674	0.000260	0.008696	0.219176
0	((, furniture_decor))	((, bed_bath_table))	0.062976	0.097800	0.000780	0.012380	0.126582
1	((, bed_bath_table))	((, furniture_decor))	0.097800	0.062976	0.000780	0.007972	0.126582
9	((, housewares))	((, garden_tools))	0.060811	0.042273	0.000260	0.004274	0.101093
8	((, garden_tools))	((, housewares))	0.042273	0.060811	0.000260	0.006148	0.101093



ASSOCIATION RULES FOR PARANA

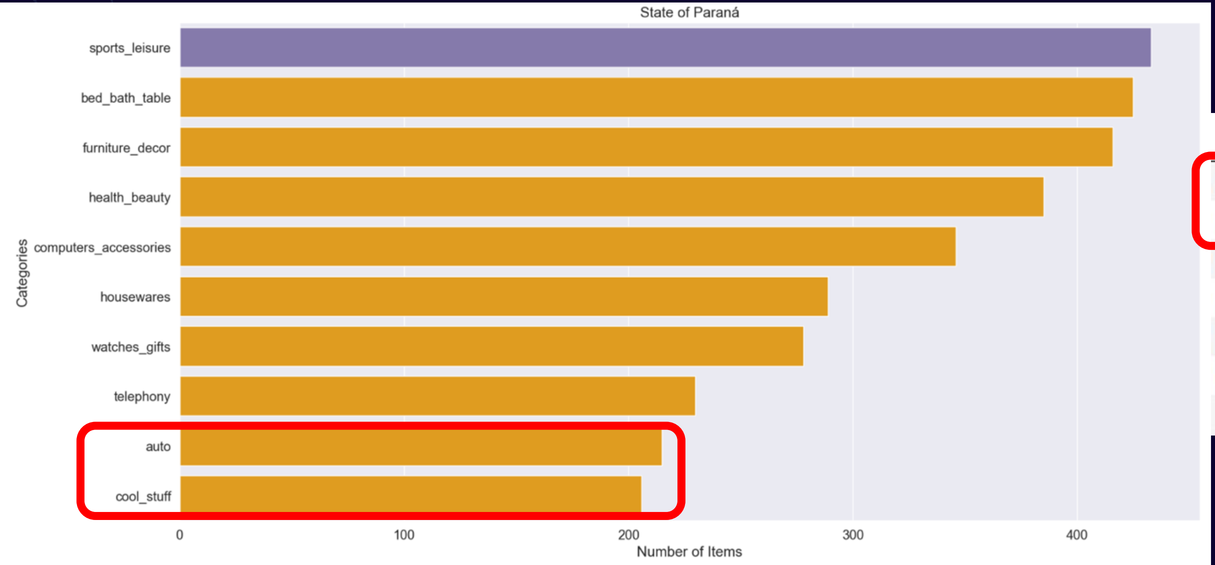
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
47	((, cool_stuff))	((, bed_bath_table), (, auto))	0.040616	0.000200	0.000200	0.004926	24.620690
42	((, bed_bath_table), (, auto))	((, cool_stuff))	0.000200	0.040616	0.000200	1.000000	24.620690
43	((, bed_bath_table), (, cool_stuff))	((, auto))	0.000200	0.041817	0.000200	1.000000	23.913876
46	((, auto))	((, bed_bath_table), (, cool_stuff))	0.041817	0.000200	0.000200	0.004785	23.913876
45	((, bed_bath_table))	((, cool_stuff), (, auto))	0.080432	0.000200	0.000200	0.002488	12.432836
44	((, cool_stuff), (, auto))	((, bed_bath_table))	0.000200	0.080432	0.000200	1.000000	12.432836
6	((, furniture_decor))	((, home_confort))	0.079432	0.002601	0.000400	0.005038	1.936834
7	((, home_confort))	((, furniture_decor))	0.002601	0.079432	0.000400	0.153846	1.936834
17	((, furniture_living_room))	((, toys))	0.002801	0.040216	0.000200	0.071429	1.776119
16	((, toys))	((, furniture_living_room))	0.040216	0.002801	0.000200	0.004975	1.776119
41	((, art))	((, furniture_decor))	0.001801	0.079432	0.000200	0.111111	1.398825
40	((, furniture_decor))	((, art))	0.079432	0.001801	0.000200	0.002519	1.398825
10	((, furniture_decor))	((, industry_commerce_and_business))	0.079432	0.002001	0.000200	0.002519	1.258942
11	((, industry_commerce_and_business))	((, furniture_decor))	0.002001	0.079432	0.000200	0.100000	1.258942
25	((, electronics))	((, food))	0.033613	0.005402	0.000200	0.005952	1.101852
24	((, food))	((, electronics))	0.005402	0.033613	0.000200	0.037037	1.101852
38	((, audio))	((, watches_gifts))	0.003401	0.054822	0.000200	0.058824	1.072993
39	((, watches_gifts))	((, audio))	0.054822	0.003401	0.000200	0.003650	1.072993

Association Rules for Rio Grande do Sul

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
55	((, construction_tools_lights))	((, bed_bath_table), (, garden_tools))	0.003314	0.000368	0.000184	0.055556	150.888889
50	((, bed_bath_table), (, garden_tools))	((, construction_tools_lights))	0.000368	0.003314	0.000184	0.500000	150.888889
54	((, garden_tools))	((, bed_bath_table), (, construction_tools_lig...	0.042710	0.000184	0.000184	0.004310	23.413793
51	((, bed_bath_table), (, construction_tools_lig...	((, garden_tools))	0.000184	0.042710	0.000184	1.000000	23.413793
92	((, housewares))	((, furniture_decor), (, watches_gifts))	0.064433	0.000184	0.000184	0.002857	15.520000
...	...	...	...	...	...	...	...
67	((, cool_stuff))	((, baby))	0.046944	0.037003	0.000184	0.003922	0.105980
17	((, garden_tools))	((, furniture_decor))	0.042710	0.081738	0.000368	0.008621	0.105468
16	((, furniture_decor))	((, garden_tools))	0.081738	0.042710	0.000368	0.004505	0.105468
28	((, stationery))	((, health_beauty))	0.025221	0.072901	0.000184	0.007299	0.100125
29	((, health_beauty))	((, stationery))	0.072901	0.025221	0.000184	0.002525	0.100125

DATA-DRIVEN INSIGHTS AND RECOMMENDATIONS

PARANA



	antecedents	consequents
47	((, cool_stuff))	((, bed_bath_table), (, auto))
42	((, bed_bath_table), (, auto))	((, cool_stuff))
43	((, bed_bath_table), (, cool_stuff))	((, auto))
46	((, auto))	((, bed_bath_table), (, cool_stuff))
45	((, bed_bath_table))	((, cool_stuff), (, auto))
44	((, cool_stuff), (, auto))	((, bed_bath_table))
6	((, furniture_decor))	((, home_confort))

ARL

EDA

Recommendations

- Recommendation system: **Collaborative Filtering**
    - Product-Product based recommendation System

Collaborative filtering can customize its recommendations to customers individual shopping behaviors, however ARL can be seen as a bigger picture approach.

Reference

https://connect.in-cosmetics.com/news-category/cosmetics-and-brazilian-consumer-behavior/

https://reliefweb.int/report/brazil/yellow-fever-brazil-24-november-2017

https://towardsdatascience.com/association-rules-2-aa9a77241654

https://pbpython.com/market-basket-analysis.html

https://www.kaggle.com/yugagrawal95/market-basket-analysis-apriori-in-python

https://medium.com/swlh/a-tutorial-about-market-basket-analysis-in-python-predictive-hacks-497dc6e06b27

https://en.wikipedia.org/wiki/Rio_de_Janeiro

https://en.wikipedia.org/wiki/S%C3%A3o_Paulo

https://en.wikipedia.org/wiki/Paran%C3%A1_(state)

THANK YOU