E-commerce

Product Recommendation and Restocking

Bo Yan & Sirish Munipalli

Problems to address

1. Product Recommendation
2. Prediction of restocking to the sellers

Problems to address - Recommendation

Why problem is interesting?

Customer:

- User experience
- Customer engagement
- Quickly and easily find desired products

Problems to address - Recommendation

What real-life application is being addressed?

Customer:

- Cannot find the exact item they truly desire
- Long and effortful searches

Problems to address - Restocking

Why problem is interesting?

Sellers:

- Meet customer orders on time
- Utilize resources
- Achieve financial targets

Problems to address - Restocking

What real-life application is being addressed?

Sellers:

- Lose customers
- Lose potential sales
- Overstock products

Dataset

E-commerce

Dataset: Source

- Canvas
- Identifiable and openly available
- 100,000 samples
- Information included
  - customers
  - orders
  - products
  - sellers
- https://canvas.ucsd.edu/courses/25097/files/4055460?module_item_id=868706

Dataset: Description

1. products_dataset.csv
2. customers_dataset.csv
3. sellers_dataset.csv
4. orders_dataset.csv
5. geolocation_dataset.csv
6. product_category_name_translation.csv
7. order_payments_dataset.csv
8. order_items_dataset.csv
9. customer_reviews_dataset.csv

Dataset: EDA Findings

1. Data quality issues/Clean:

- Missing values (drop)
- Duplicated rows (drop)
- Unused columns (drop)
- Unmatched Data type (convert)
- Others

Dataset: EDA Findings

2. Summary statistics:

	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_length_cm	product_height_cm	product_width_cm
count	32327.000000	32327.000000	32327.000000	32327.000000	32327.000000	32327.000000	32327.000000
mean	48.473722	771.517277	2.188790	2276.960807	30.856498	16.955950	23.208464
std	10.246346	635.189674	1.736767	4279.734063	16.958460	13.637246	12.080665
min	5.000000	4.000000	1.000000	0.000000	7.000000	2.000000	6.000000
25%	42.000000	339.000000	1.000000	300.000000	18.000000	8.000000	15.000000
50%	51.000000	595.000000	1.000000	700.000000	25.000000	13.000000	20.000000
75%	57.000000	972.000000	3.000000	1900.000000	38.000000	20.500000	30.000000
max	76.000000	3992.000000	20.000000	40425.000000	105.000000	105.000000	118.000000

	payment_sequential	payment_installments	payment_value
count	96460.000000	96460.000000	96460.000000
mean	1.023170	2.911829	157.202793
std	0.224815	2.704617	216.829789
min	1.000000	0.000000	0.000000
25%	1.000000	1.000000	60.000000
50%	1.000000	2.000000	103.130000
75%	1.000000	4.000000	174.410000
max	17.000000	24.000000	13664.080000

	survey_score
count	100000.000000
mean	4.070890
std	1.359663
min	1.000000
25%	4.000000
50%	5.000000
75%	5.000000
max	5.000000

Dataset: EDA Findings

3. Common features:

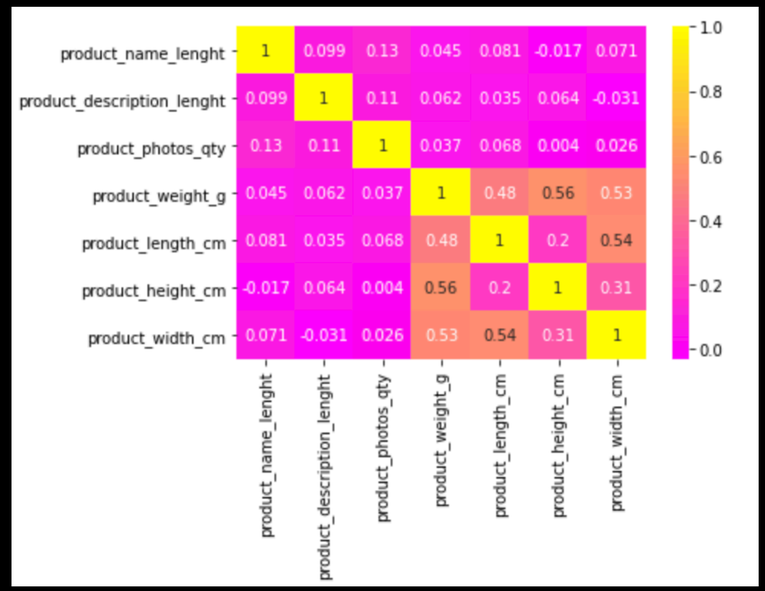
Merge: Inner join, Left on, Right on

- orders_dataset.order_it
- order_payments_dataset.order_id
- order_items_dataset.order_id
- order_items_dataset.product_id
- products_dataset.product_id

Dataset: EDA Findings

4. Variable relationships

- **Strong correlation:**
  - product_height_cm & product_weight_g
  - product_width_cm & product_weight_g
  - product_length_cm & product_width_cm
- **Medium correlation:**
  - product_length_cm & product_weight_g
  - product_height_cm & product_width_cm
  - payment_installments & payment_value,
  - price & freight_value
  - geo_lat & geo_lng
- **Weak correlation:**
  - product_name_length & product_weight_g
  - product_name_length & product_length_cm
  - product_name_length & product_height_cm
  - product_name_length & product_width_cm



Dataset: EDA Findings

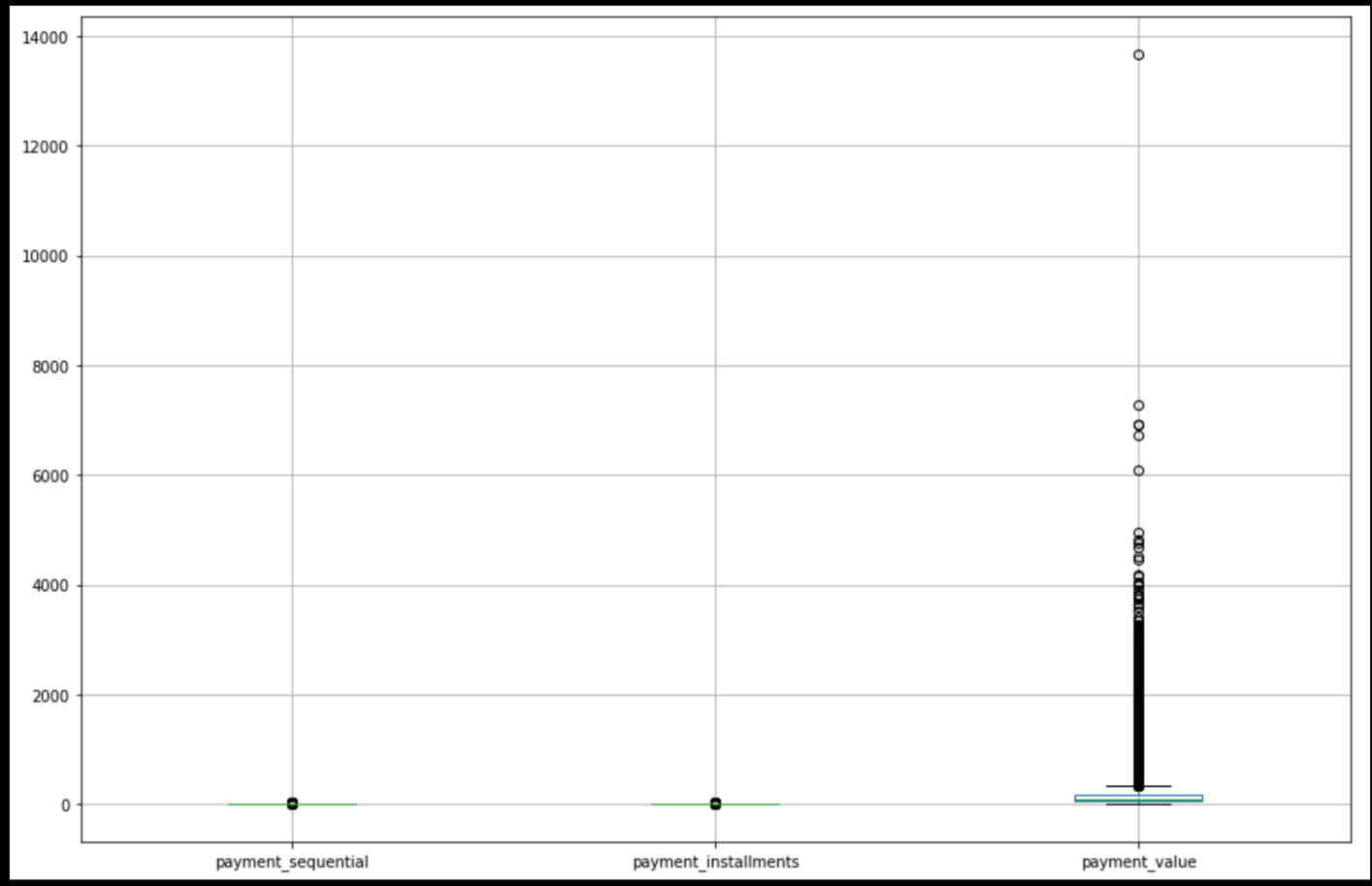
5. Data distribution:

- Highly skewed distribution:
  - product_weight_g
  - payment_value
  - price
  - freight_value
  - survey_score
- Moderately skewed distribution:
  - product_name_lenght
- Approximately symmetric distribution:
  - none

Dataset: EDA Findings

6. Outliers:

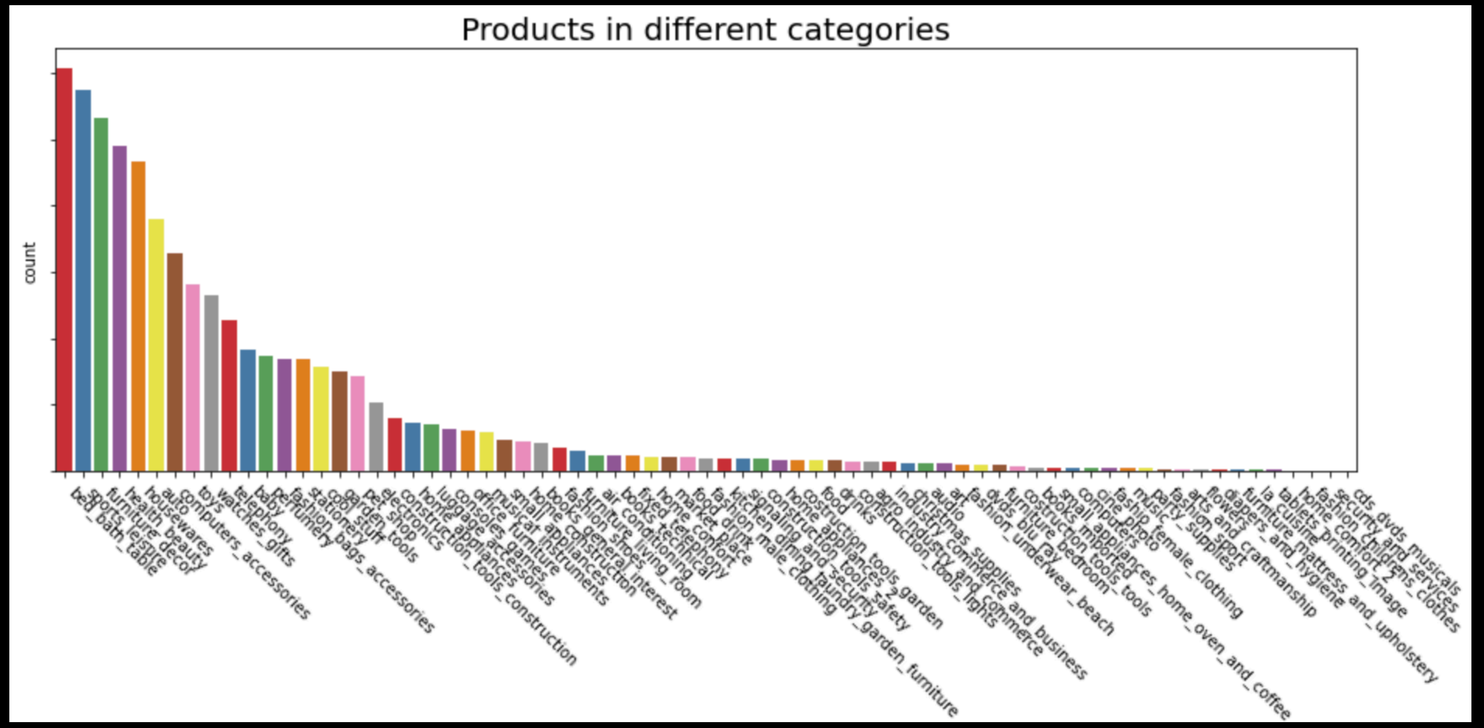
- product_weight_g
- payment_value
- price
- survey_score



Dataset: EDA Findings

7. Common trends:

- Most of product weight
  - (0, 10,000) gram
- Most products belong to
  - bed_bath_table category
- Most customers live in
  - SP state
- Most popular payment type
  - credit_card
- Most products' survey score
  - 5



Analysis task planned

- **Type of tasks:**
  - Unsupervised Learning: Clustering
  - Supervised Learning: Regression
- **How does task related to business problem:**
  - Assign each cluster to customer preferences
  - Product features and geolocation to do cluster
  - Estimates relationship between multiple variables
  - Predict future sales of the products based on the past information

Insights to gain

- **Recommendation:**
  - Clusters make intuitive sense
- **Restocking:**
  - Identify factors matter most
  - High accuracy in prediction

Potential challenges

- **Data challenges:**
  - Volume
  - Velocity
  - Effects of Outliers
- **Task challenges:**
  - Google API to get the geolocation data
  - Identify most influential features

Thank you!

Q & A