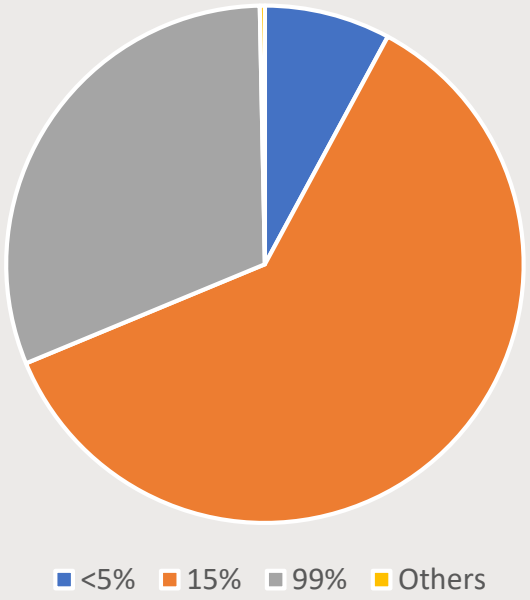
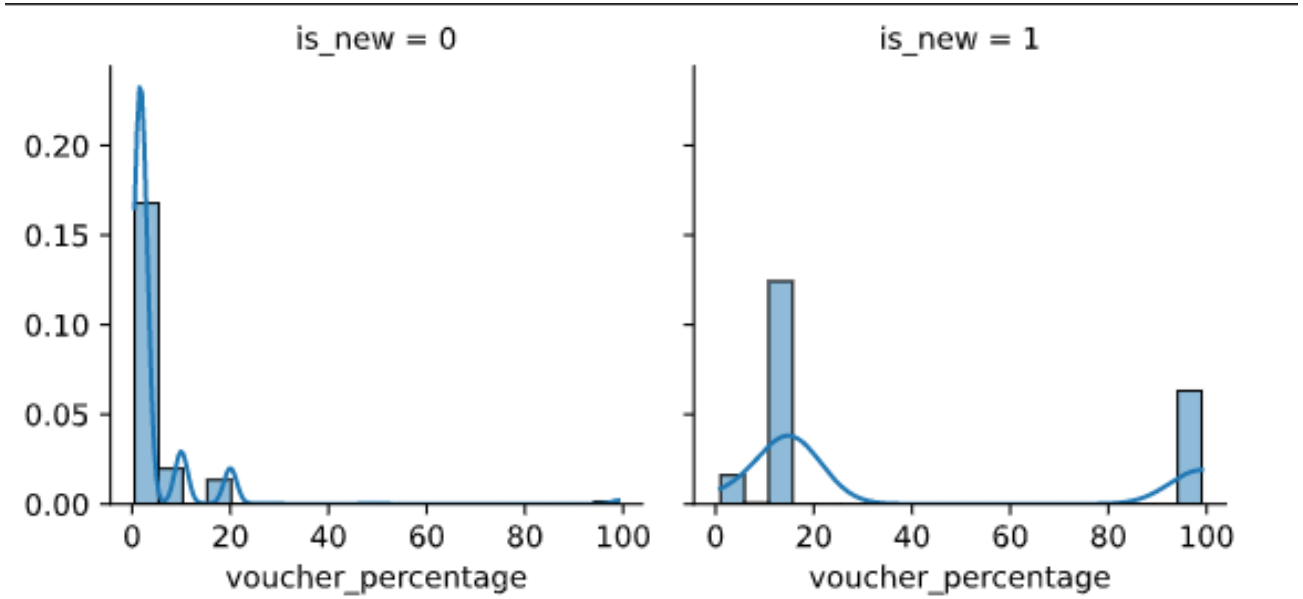
Studies

- `new_user`
  - More user = more potential for revenue
- `is_paid`
  - Completed transaction is a revenue
- `gmv` (Gross Merchandise Value)
  - More GMV = more income
- `basket_amount`

new_user
Voucher percentage

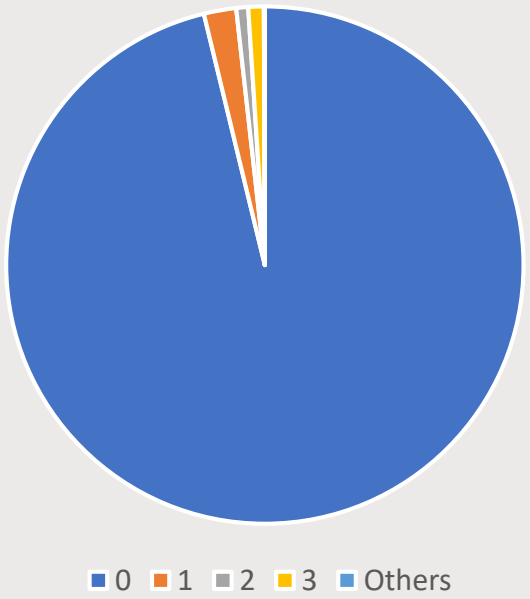
- Zero voucher is removed
- Voucher correlates to new user joining
- Vouchers mainly used

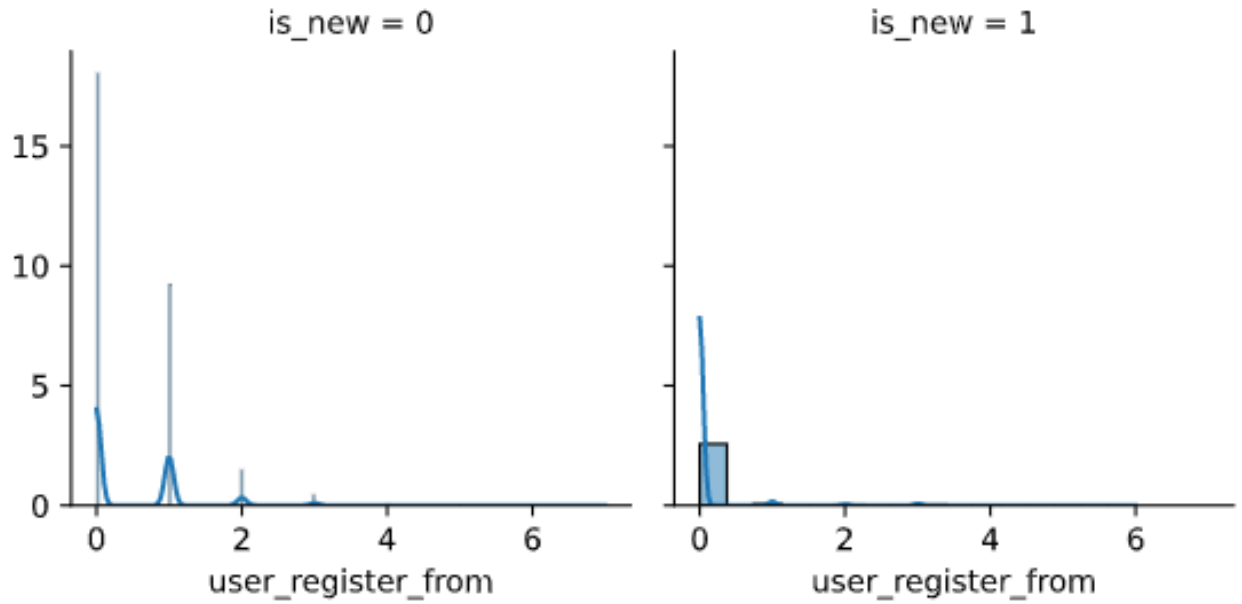




new_user
User registered from

- New user mainly come from platform 0



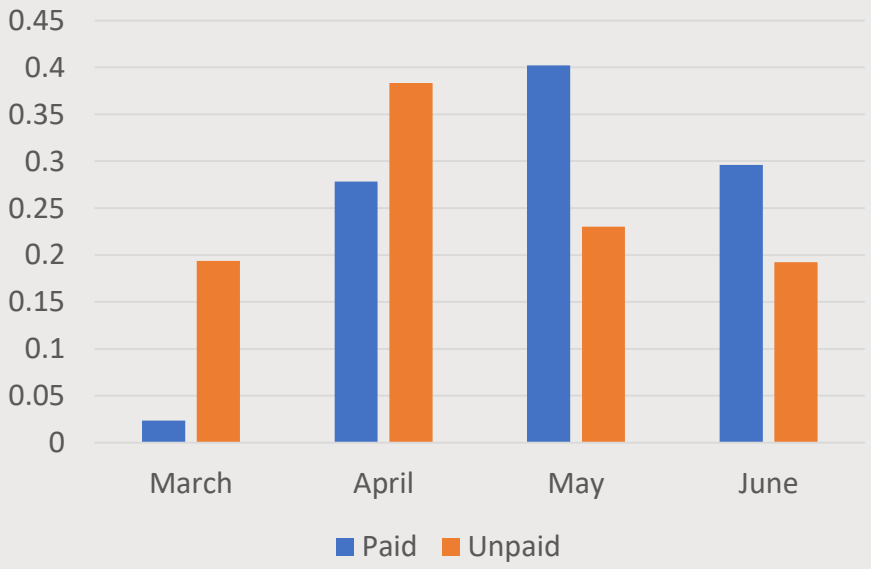


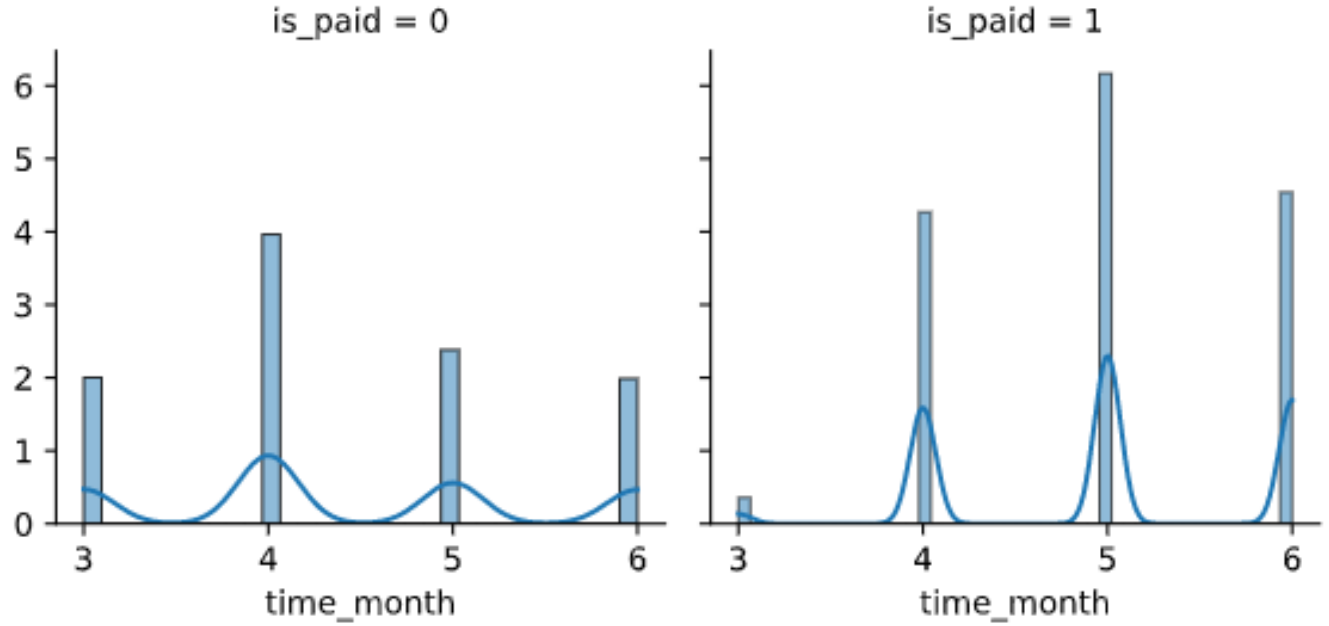
`new_user` – Other Findings

- New user are likely to use valid vouchers
- New user are likely to do transaction with vouchers

is_paid
Voucher percentage

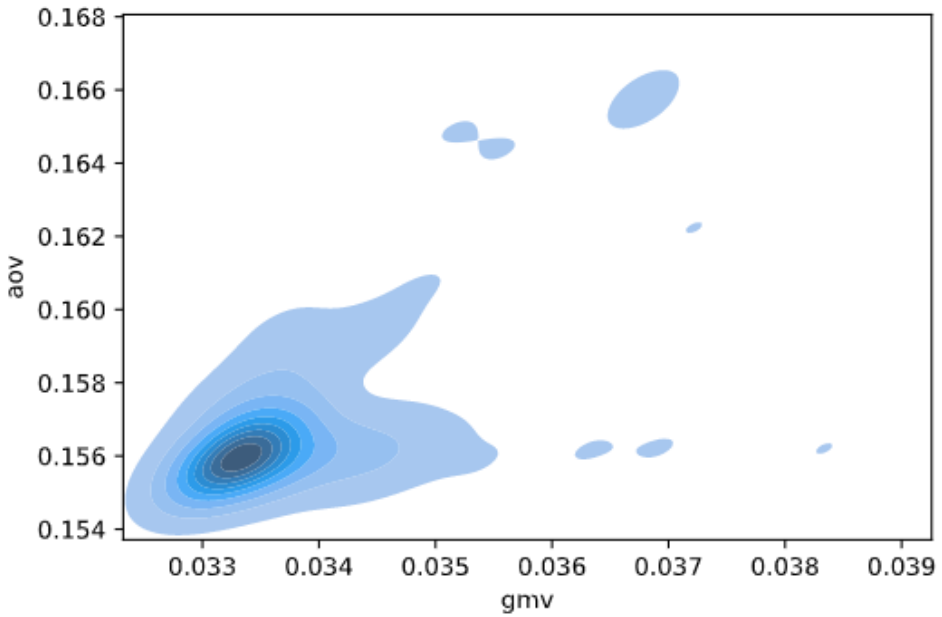
- May and June
  - More completed transaction
- March
  - More incomplete transaction

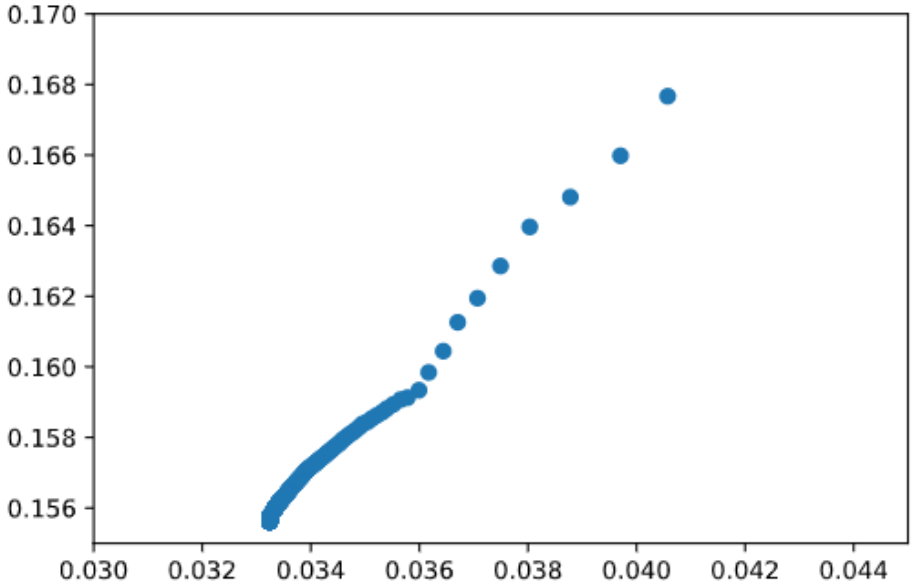




gmv
Average Order Value

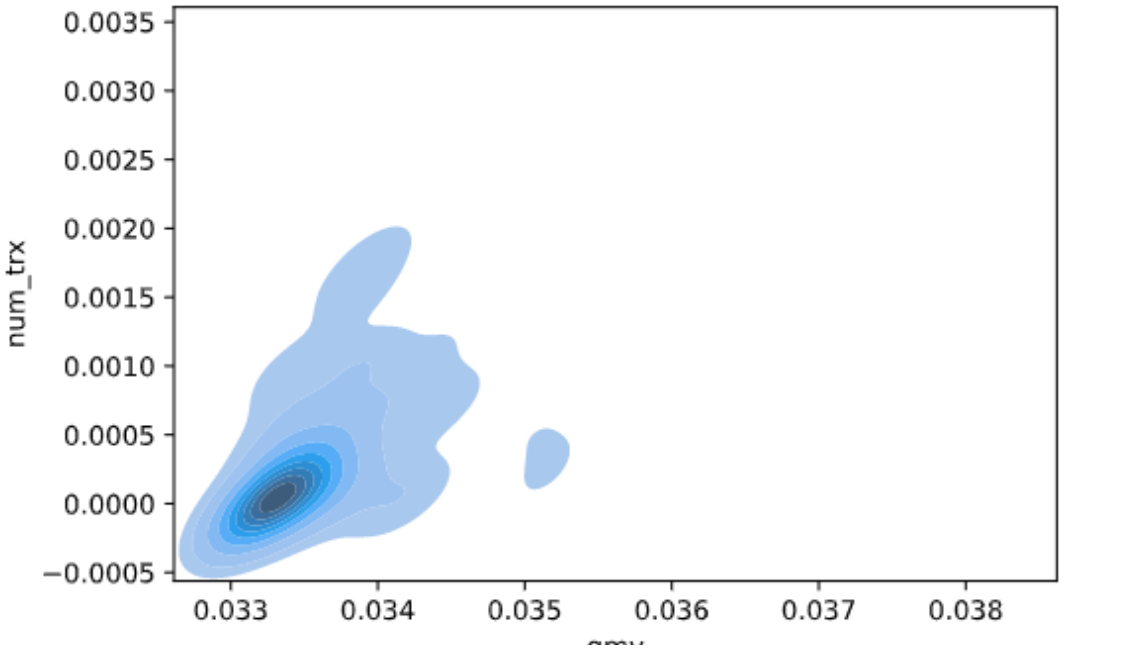
- Linear Relationship
- $aov \propto gmv$
- More aov = more gmv

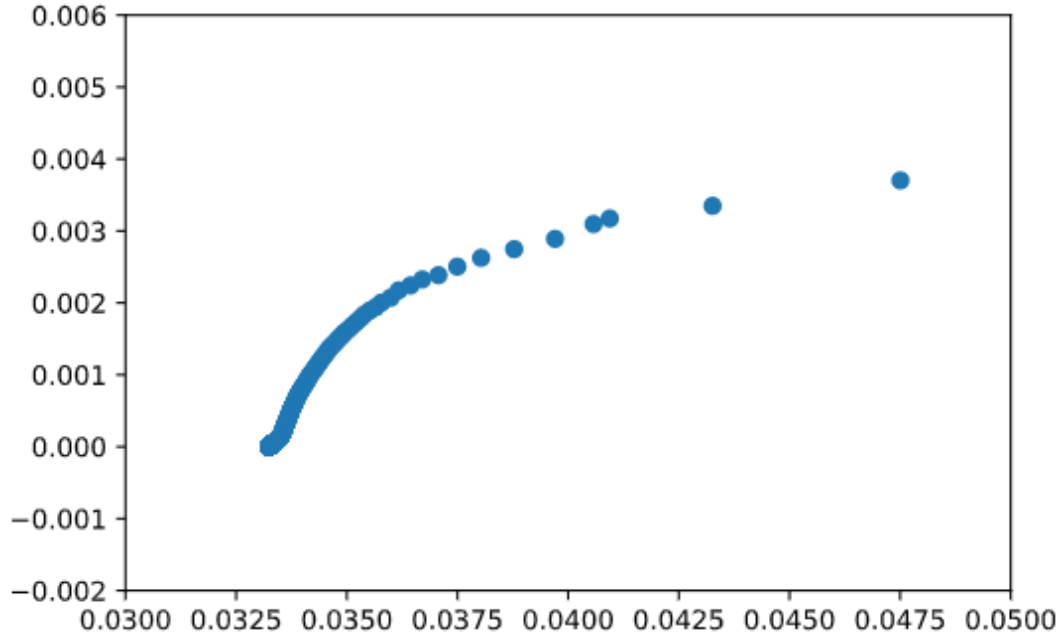




gmv
Number of transaction

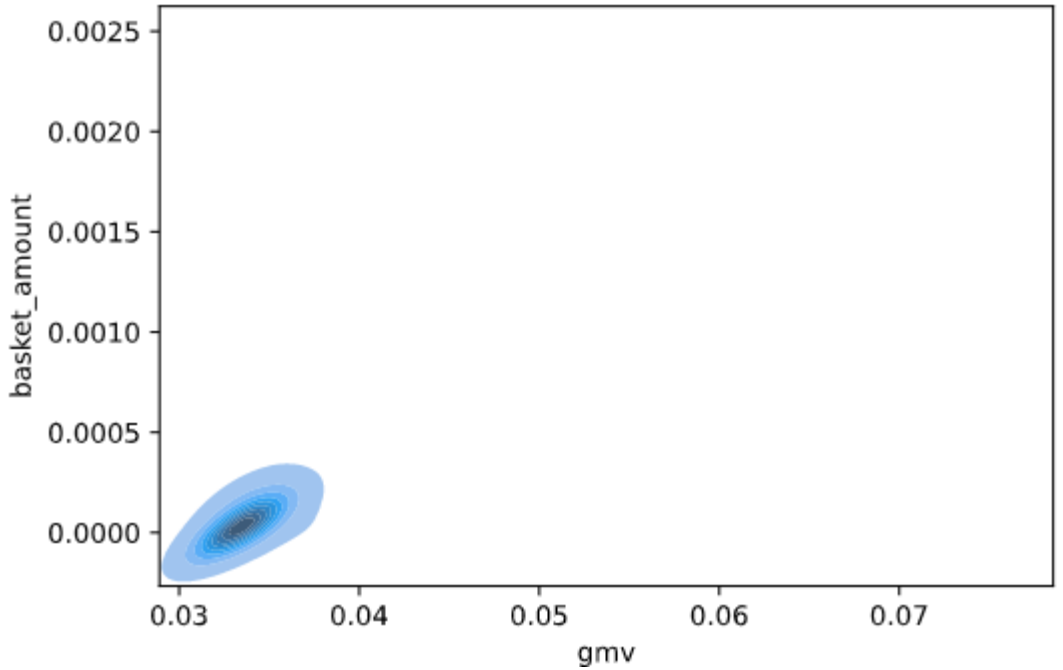
- Squared Relationship
- $n_{trx}^2 \propto gmv$
- More transaction = more gmv

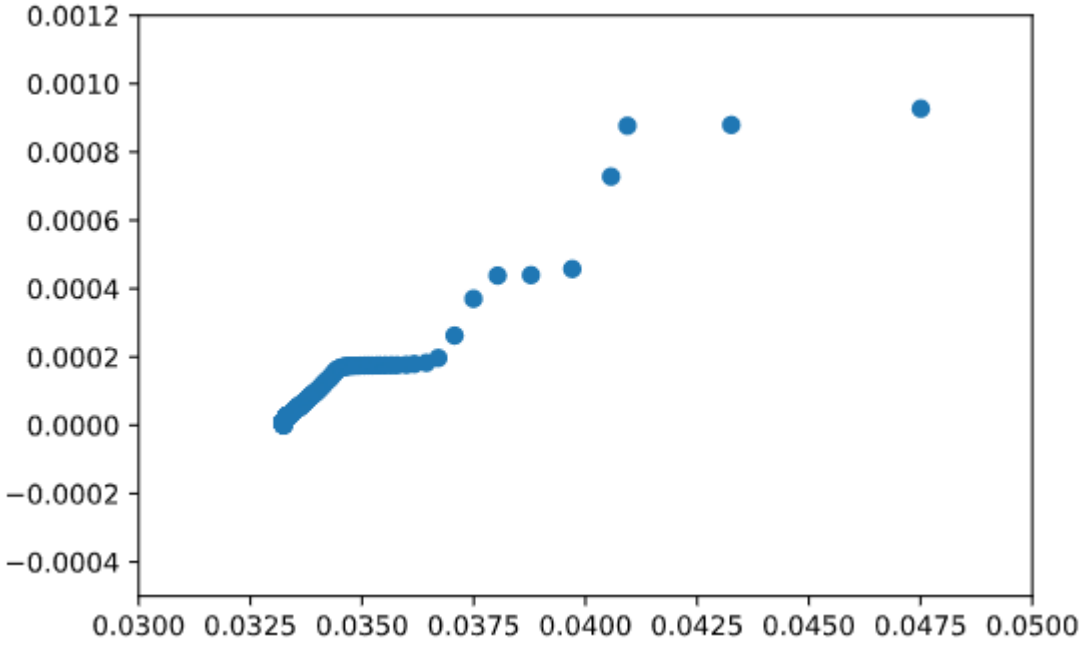




gmv
Basket Amount

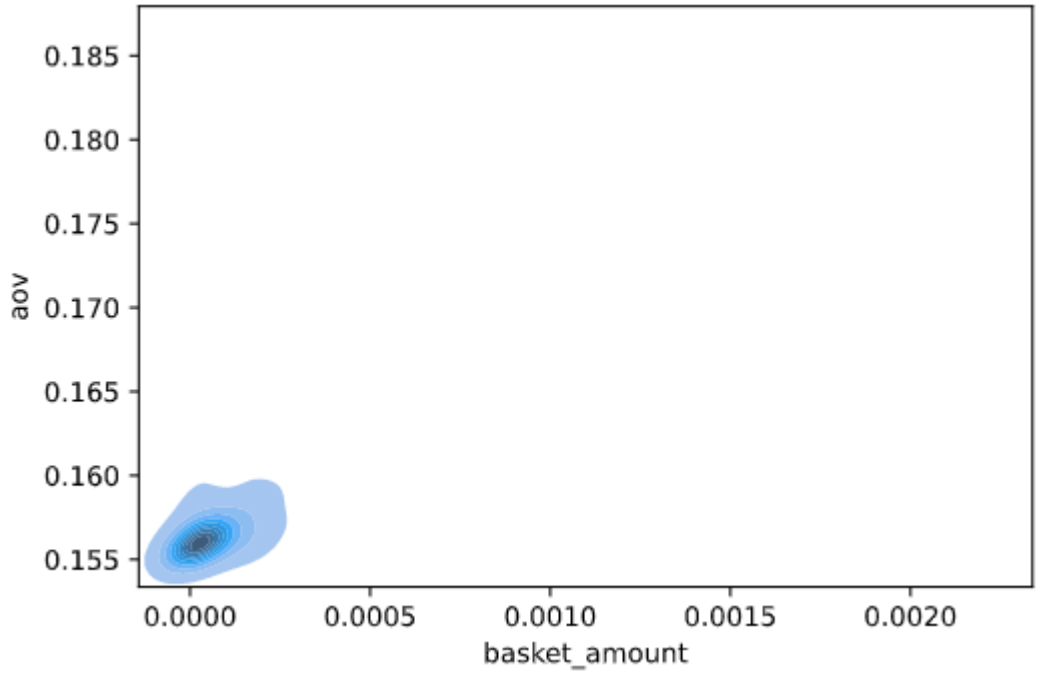
- Linear Relationship
- $basketAmount \propto gmv$
- More basket amount = more gmv

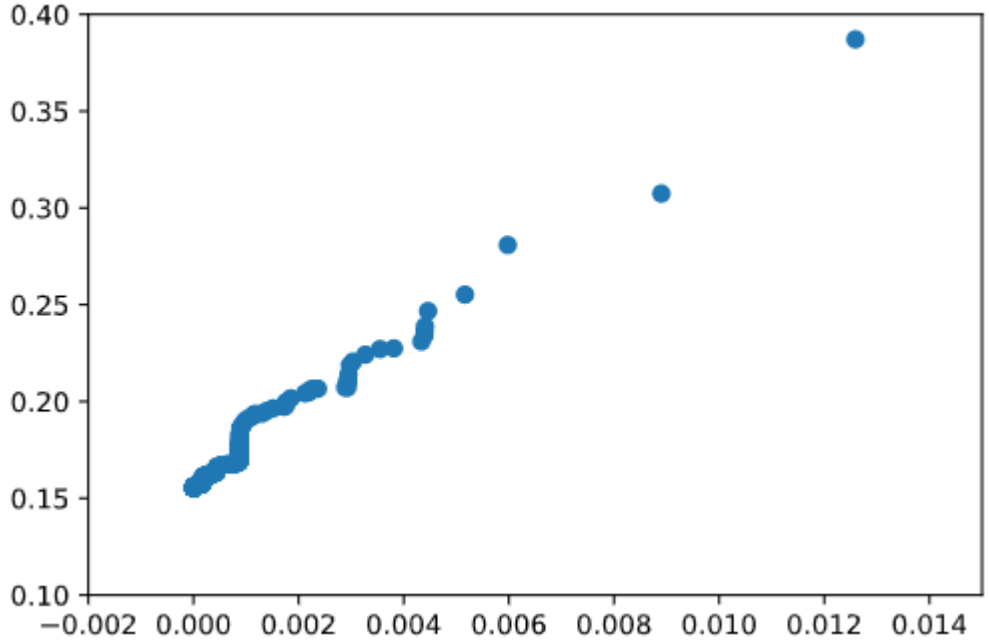




basket_amount
Average order Value

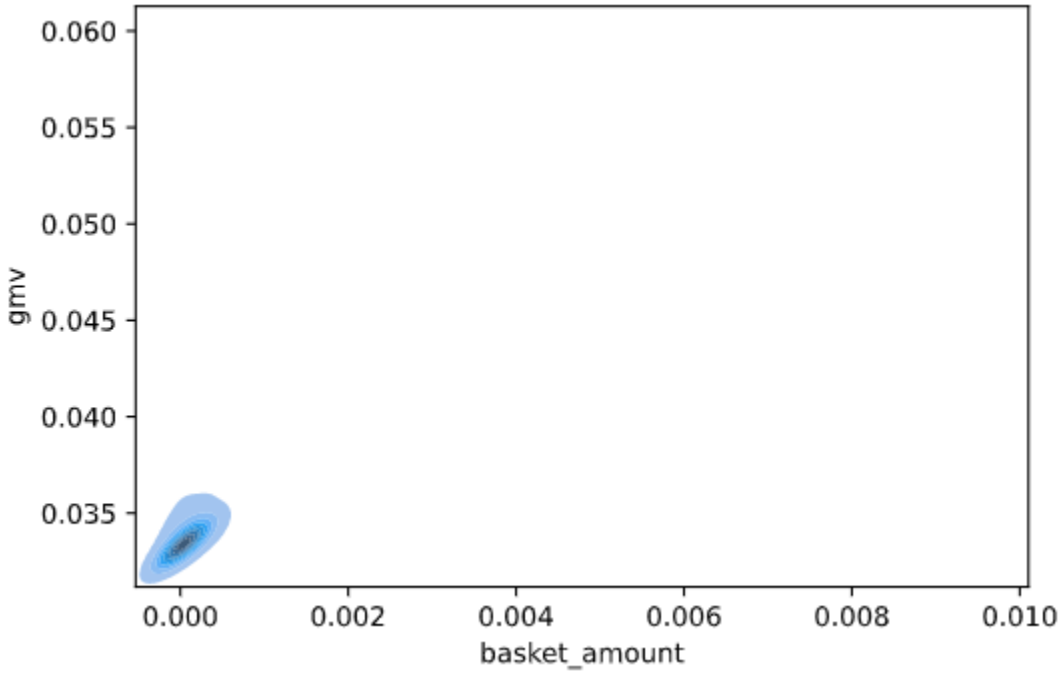
- Linear Relationship
- $basketAmount \propto aov$
- More average order volume = more aov (obvious relationship)

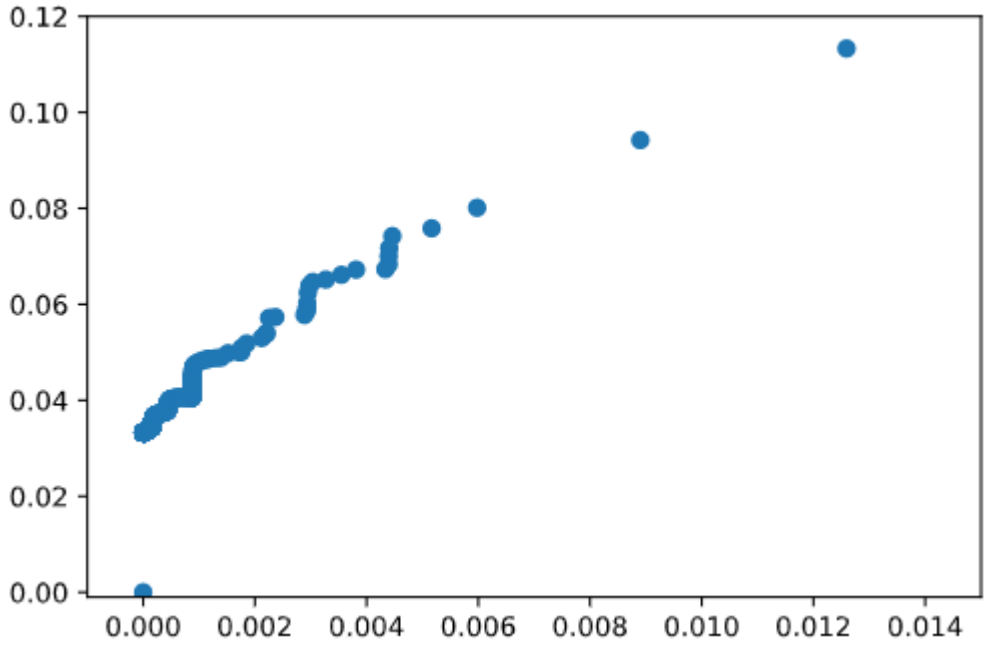




basket_amount
gmv

- Linear Relationship
- $basketAmount \propto gmv$
- More basket amount = more gmv





`basket_amount` – Other Findings

- Higher basket amount will unlikely to be purchased nor paid

Machine Learning

- Outliners
  - Not removed
  - Lowers accuracy
- Normalisation
  - Feature scaling (0-1)
- Most data don’t follow normal distribution
  - Poor normal-probability plot
  - Poor kurtosis value
  - Good Q-Q plot
- 10% testing, 90% training
  - 5 Kfold (18% validation, 72% training)

Machine Learning

- Large data = Slow performance
  - Random Under-Sampling (fast)
- Model can be built into API
  - Python backend (i.e. Flask or Django)
  - Easy implementation
    - Get request / REST API
    - Dockerised
    - AWS ECS

`new_user`

- Classification
- Data is unevenly distributed (87% vs 13%)
  - Oversampling
    - Increases false positive for new user
    - Decreases false positive for old user
    - Uneven = don’t use
- Baseline model = Logistic Regression (95% paper accuracy)
  - 78% Accuracy for `new_user = 1`
- Best model = XGBoost (98% paper accuracy)
  - Also the fastest due to GPU support

new_user
Feature importance

- Voucher Valid
- Account type
- Referrer type
- Possible Features:
  - Voucher percentage
  - Voucher amount
- Confirms data exploration

XGBoost

Weight	Feature
0.2716	user_purchased_prior
0.2257	account_age
0.0893	gmv
0.0496	user_group
0.0467	voucher_percentage
0.0418	voucher_valid
0.0401	voucher_type
0.0358	aov
0.0285	user_register_from
0.0191	referrer_type
0.0161	voucher_max_amount
0.0143	voucher_min_purchase
0.0105	num_voucher_errors
0.0105	sessions
0.0095	user_type
0.0086	voucher_amount
0.0084	average_session_length
0.0080	account_type
0.0070	time_day
0.0069	purchase
	… 11 more …

Logistic Regression

Weight[?]	Feature
+5.081	voucher_valid
+1.851	account_type
+1.172	purchase
+0.633	<BIAS>
+0.499	referrer_type
+0.307	voucher_amount
+0.246	voucher_min_purchase
+0.235	marketing_tier
	… 4 more positive …
	… 8 more negative …
-0.344	num_voucher_errors
-0.414	user_register_from
-0.418	trx_is_voucher
-0.466	aov
-0.882	is_paid
-1.114	user_group
-3.616	num_visit_promo_page
-5.074	sessions
-6.209	voucher_type
-7.236	user_purchased_prior
-7.866	average_session_length
-12.200	num_product_types

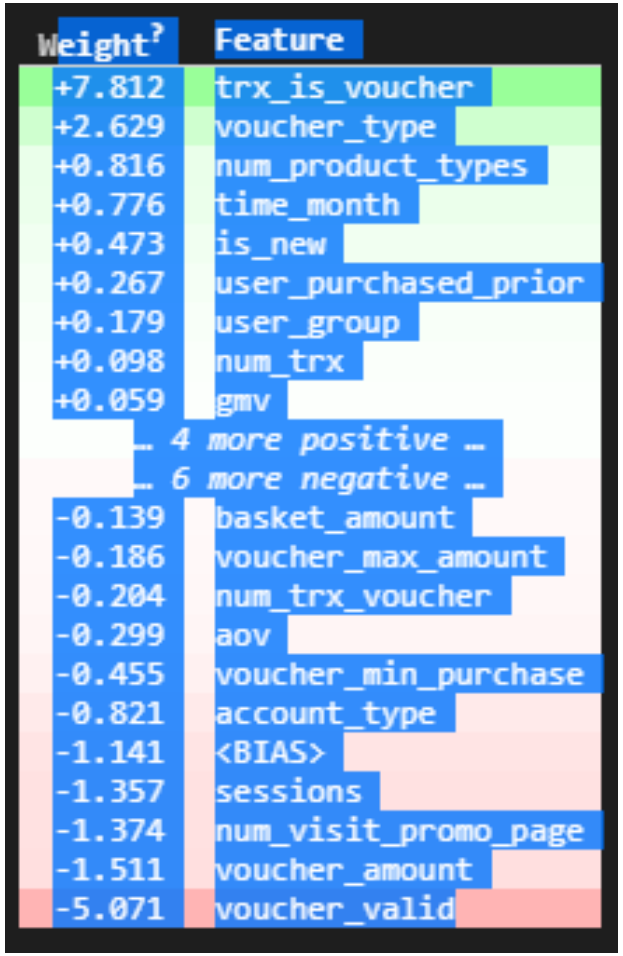
`is_paid`

- Classification
- Data is unevenly distributed (96% vs 4%)
- Baseline model = Logistic Regression (96% paper accuracy)
  - 95% Accuracy for `is_paid = 0`
  - Best model

is_paid
Feature importance

- Transaction using voucher
- Voucher type
- Number of product types
- Time of the month
- Possible Features:
  - User group
- Negative (opposite):
  - Voucher needs to be valid
  - High voucher amount
  - Visiting promo page
  - Sessions

Logistic Regression



Weight	Feature
+7.812	trx_is_voucher
+2.629	voucher_type
+0.816	num_product_types
+0.776	time_month
+0.473	is_new
+0.267	user_purchased_prior
+0.179	user_group
+0.098	num_trx
+0.059	gmv
… 4 more positive …	
… 6 more negative …	
-0.139	basket_amount
-0.186	voucher_max_amount
-0.204	num_trx_voucher
-0.299	aov
-0.455	voucher_min_purchase
-0.821	account_type
-1.141	<BIAS>
-1.357	sessions
-1.374	num_visit_promo_page
-1.511	voucher_amount
-5.071	voucher_valid

gmv

- Regression
- Normalisation
  - Optional
  - Feature Selection
- Safe model = Linear Regression (98.85%)
- Best model = Decision Tree (99.98%)
  - Possible overfitting on outside of dataset

gmv
Feature importance

- aov
- Number transaction
- Possible features:
  - Province
  - Basket amount
  - Visiting promo page
  - Number of transaction voucher
- Confirms the data exploration

Decision Tree

Weight	Feature
0.6658	aov
0.3286	num_trx
0.0029	province
0.0011	num_visit_promo_page
0.0003	num_trx_voucher
0.0003	account_age
0.0002	user_type
0.0001	average_session_length
0.0001	voucher_percentage
0.0001	referrer_type
0.0001	marketing_tier
0.0001	sessions
0.0000	num_product_types
0.0000	time_day
0.0000	user_group
0.0000	num_voucher_errors
0.0000	account_type
0.0000	user_register_from
0.0000	voucher_amount
0.0000	basket_amount
	… 11 more …

Linear Regression

Weight[?]	Feature
+0.936	num_trx
+0.492	aov
+0.050	basket_amount
+0.011	num_trx_voucher
+0.004	voucher_min_purchase
+0.002	num_visit_promo_page
+0.001	num_product_types
+0.000	is_new
+0.000	voucher_valid
+0.000	is_paid
+0.000	user_group
	… 9 more positive …
	… 3 more negative …
-0.000	trx_is_voucher
-0.000	purchase
-0.000	average_session_length
-0.000	voucher_type
-0.001	account_type
-0.001	sessions
-0.004	voucher_max_amount
-0.012	voucher_amount
-0.044	<BIAS>

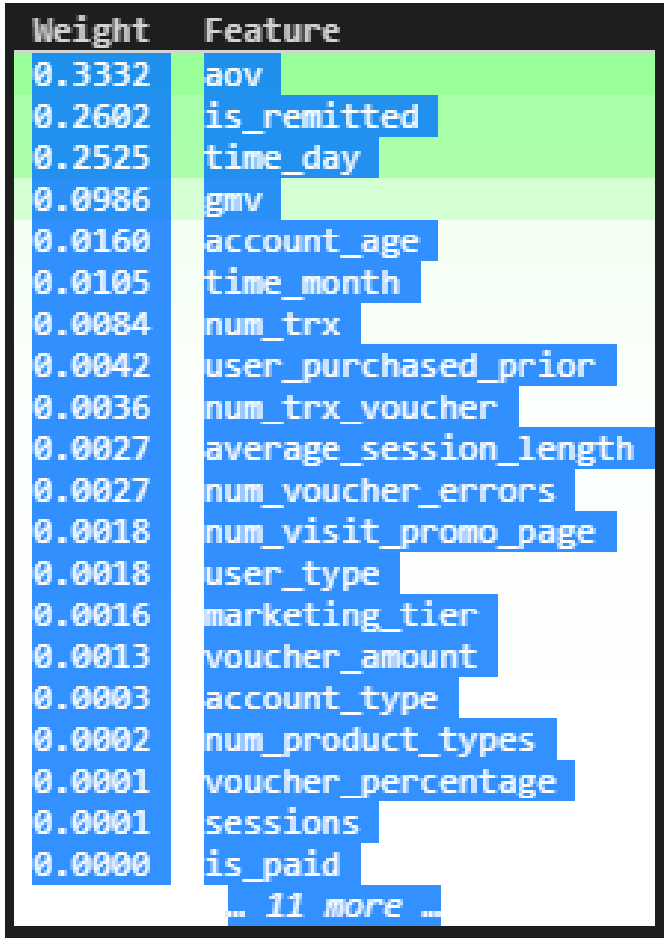
basket_amount

- Regression
- Normalisation
  - Standatisation
- Safe model = Linear Regression (96.69%)
- Best model = Decision Tree (98.38%)
  - Possible overfitting on outside of dataset

basket_amount Feature importance

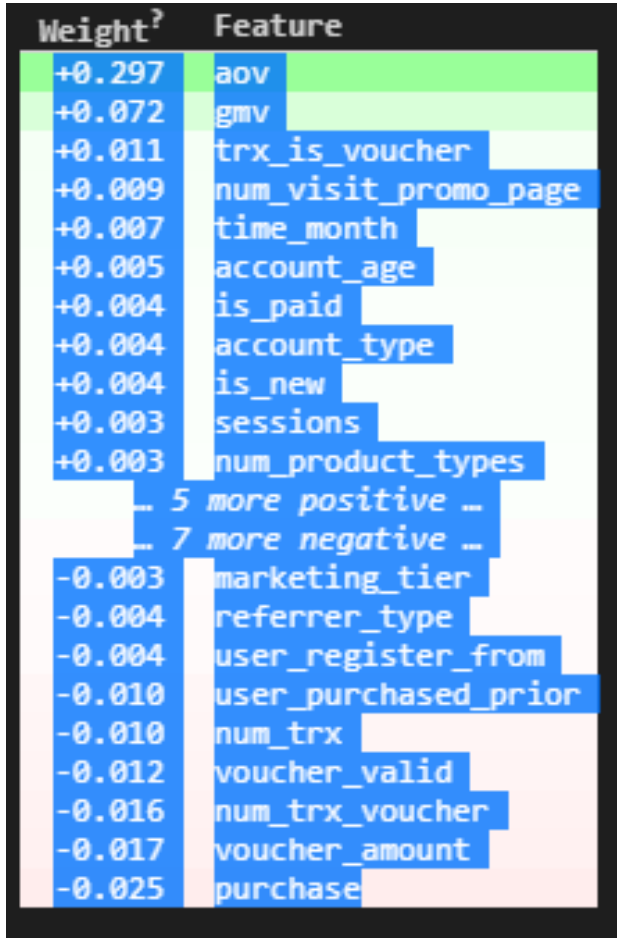
- aov
- Gmv
- is_remitted
- Days of the month
- Possible features:
  - Account age
  - Month of the year
  - Number of transaction
  - User purchased before
- Confirms the data exploration

Decision Tree



Weight	Feature
0.3332	aov
0.2602	is_remitted
0.2525	time_day
0.0986	gmv
0.0160	account_age
0.0105	time_month
0.0084	num_trx
0.0042	user_purchased_prior
0.0036	num_trx_voucher
0.0027	average_session_length
0.0027	num_voucher_errors
0.0018	num_visit_promo_page
0.0018	user_type
0.0016	marketing_tier
0.0013	voucher_amount
0.0003	account_type
0.0002	num_product_types
0.0001	voucher_percentage
0.0001	sessions
0.0000	is_paid
	… 11 more …

Linear Regression



Weight?	Feature
+0.297	aov
+0.072	gmv
+0.011	trx_is_voucher
+0.009	num_visit_promo_page
+0.007	time_month
+0.005	account_age
+0.004	is_paid
+0.004	account_type
+0.004	is_new
+0.003	sessions
+0.003	num_product_types
	… 5 more positive …
	… 7 more negative …
-0.003	marketing_tier
-0.004	referrer_type
-0.004	user_register_from
-0.010	user_purchased_prior
-0.010	num_trx
-0.012	voucher_valid
-0.016	num_trx_voucher
-0.017	voucher_amount
-0.025	purchase

Correlation ≠ Causation

- Need more study
- Need to find the context of the data
- A/B testing

Action Plans – Short term

- Allow more voucher validity
  - Deadline, Usage restrictions
  - However, min/max amount doesn’t matter

Action Plans – Long term

- Hypothesis
  - Voucher amount doesn’t matter for new users
  - Promo page is repulsive to new user
  - Expensive item gives more revenue
    - Quality > Quantity
  - Sale follows a seasonal monthly cycle of years
    - Sale follows a seasonal daily cycle of months
  - Users with high transaction will buy more expensive item
  - High session doesn’t increase completed transaction