Buka Purchase Insights

Rizky Luthfianto

Table of contents

Research Questions

- What are at least 2 factors contribute to users making a purchase?
  - How are the calculation of their potential impact?
- What actions we can take to improve conversion?
- How much impact removing vouchers would have on the business?
  - on basket amount
  - on unique buyers

Definition - Customer Journey

=> users‘ voucher could be either valid/invalid and transaction could be with or without voucher, then

=> purchase, then

=> is_paid, then

=> is_remitted

Exploratory Data Analysis (EDA)

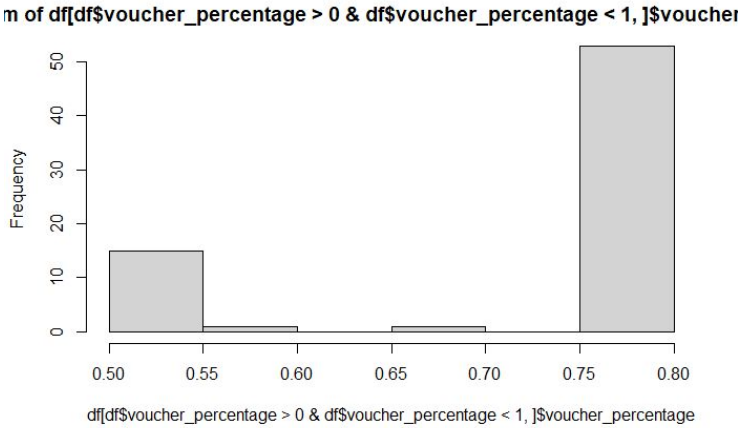
Columns summary

Before any preprocessing/imputation

	has nan?	dtypes	num of unique values	list of unique values
user_id	False	int64	69306	[281605921, 125302602, 125327377, 125601271, 1...
is_new	False	int64	2	[1, 0]
time	False	object	229835	[2015-06-12 03:41:44.263000+00:00, 2015-06-15 ...
voucher_type	True	float64	2	[nan, 1.0, 0.0]
voucher_valid	True	float64	2	[nan, 1.0, 0.0]
basket_amount	False	float64	125094	[2.8772007552981635e-05, 2.351516046569688e-05...
voucher_max_amount	False	float64	26	[0.0, 0.0002, 0.0004, 0.001, 2e-05, 4e-05, 0.0...
voucher_percentage	False	float64	20	[0.0, 1.0, 2.0, 1.5, 10.0, 20.0, 99.0, 50.0, 1...
voucher_min_purchase	False	float64	16	[0.0, 0.01, 0.02, 0.006, 0.002, 0.004, 0.001, ...
voucher_amount	False	float64	7451	[0.0, 0.0198636151779623, 0.0193235168800721, ...
trx_is_voucher	False	int64	2	[0, 1]
is_paid	False	int64	2	[0, 1]
is_remitted	False	int64	2	[0, 1]
user_purchased_prior	False	int64	2	[0, 1]
num_voucher_errors	False	int64	102	[0, 6, 5, 1, 3, 16, 48, 23, 4, 24, 22, 20, 19,...
purchase	False	int64	2	[0, 1]
province	False	int64	27	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...
marketing_tier	True	object	4	[tier_2, tier_1, tier_3, tier_4, nan]
user_type	False	int64	16	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...
user_group	False	int64	2	[0, 1]
account_type	False	int64	7	[0, 1, 2, 3, 4, 5, 6]
referrer_type	False	int64	7	[0, 1, 2, 3, 4, 5, 6]
account_created_at	False	object	69159	[2014-03-23 04:01:30+00:00, 2013-07-18 07:24:4...
user_register_from	False	int64	8	[0, 1, 2, 3, 4, 5, 6, 7]
sessions	False	float64	1513	[0.0041576283441793, 0.0164497469269703, 0.024...
average_session_length	False	float64	72918	[0.0014201489039182, 0.0048480223662408, 0.003...
num_visit_promo_page	False	float64	95	[0.0, 0.0067567567567567, 0.0202702702702702, ...
num_product_types	False	float64	31	[0.0, 0.2258064516129032, 0.032258064516129, 0...
num_trx	False	float64	1868	[0.0, 9.000090000900009e-05, 7.239202826810877...
num_trx_voucher	False	float64	122	[0.0, 0.0161662817551963, 0.023094688221709, 0...
gmv	False	float64	65720	[0.0332419998003988, 0.0338819280218784, 0.033...
aov	False	float64	65714	[0.1556185124338982, 0.1577025115338126, 0.156...

EDA - Anomaly (1)

- Voucher % has “0-1 range” & “0-100 range”



- 1 user is_remitted before even paid
  - ```
    df.query('is_paid==0 & is_remitted==1')
    ```

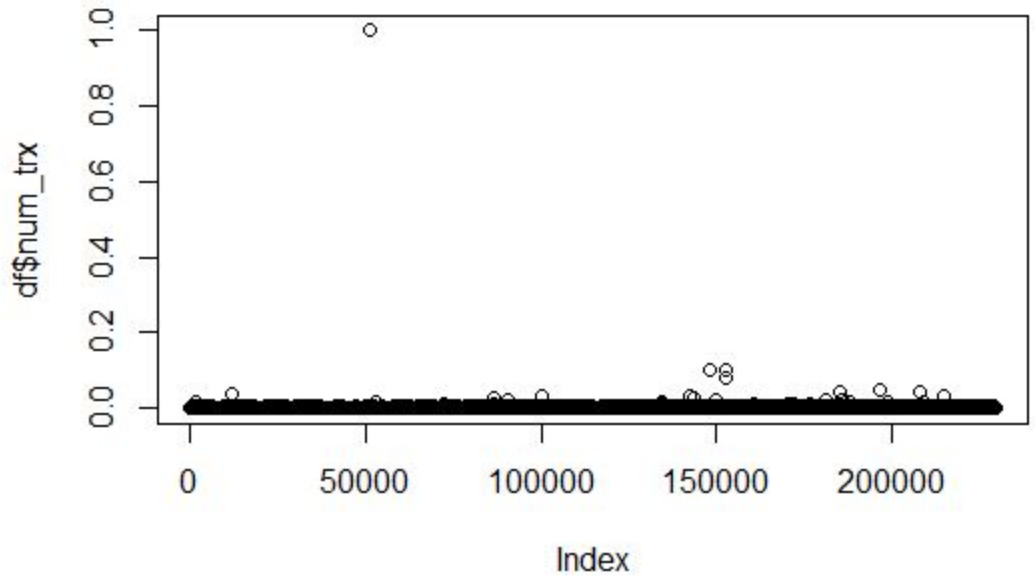
	user_id	is_new	time
101021	140386015	1	2015-06-14 10:25:21.006000+00:00

1 rows × 32 columns

- Number of voucher failure are not scaled 0..1, unlike the others
  - ```
    df$num_voucher_errors
     [1]  0  0  6  6  0  6  6  5  6  5  0  0  0  0  0  0  0  0  0  0  0  0
    [42]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
    [83]  0 23 16  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
    ```

EDA - Anomaly (2)

- num_trx==1 is outlier compared to other normalized values

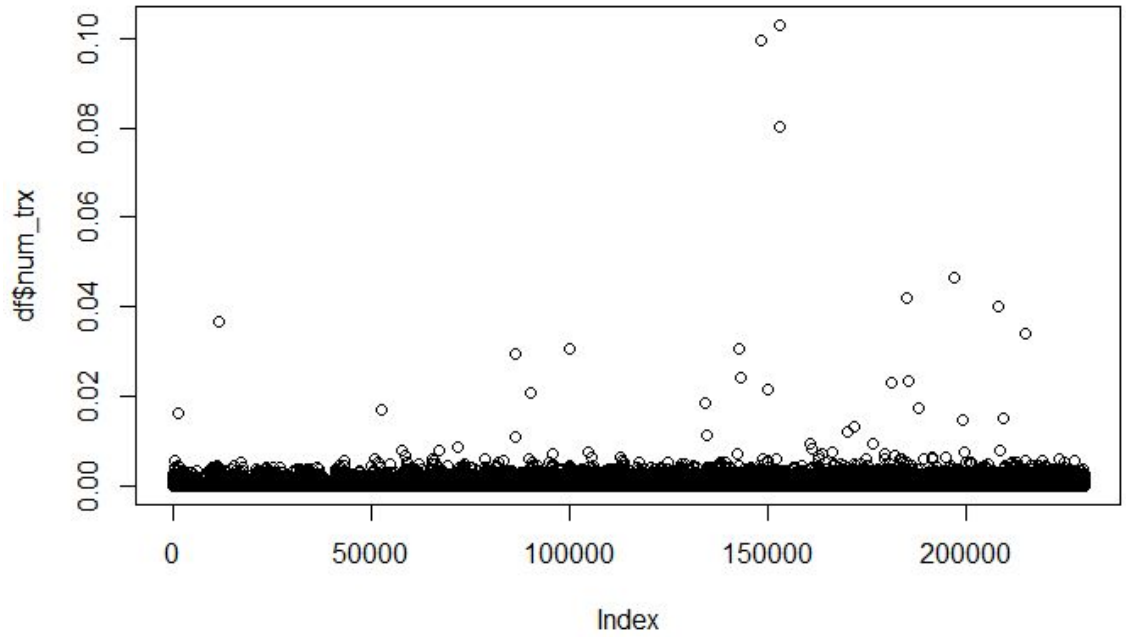


EDA - Anomaly - Action (1)

- Percentage has both “0-1 range” & “0-100 range”
  - solution: 0-1 x 100
- 1 user is_remitted before even paid
  - solution: Delete the user
- Number of voucher failure are not scaled 0..1, unlike the others
  - solution: Scaling is required if the regression is using penalty, otherwise actually is not required
- num_trx==1 is outlier compared to other normalized values
  - solution: Delete the user with num_trx==1

EDA - Anomaly - Action (2)

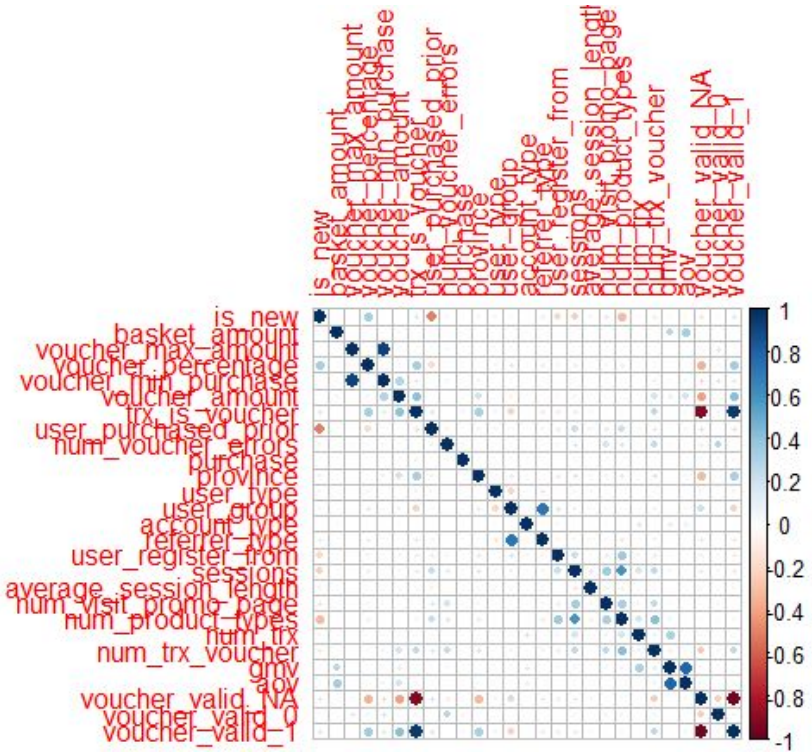
- After deleting the user with num_trx==1



EDA - missing data & correlation

- Missing value on voucher_valid
- Missing value on voucher_type

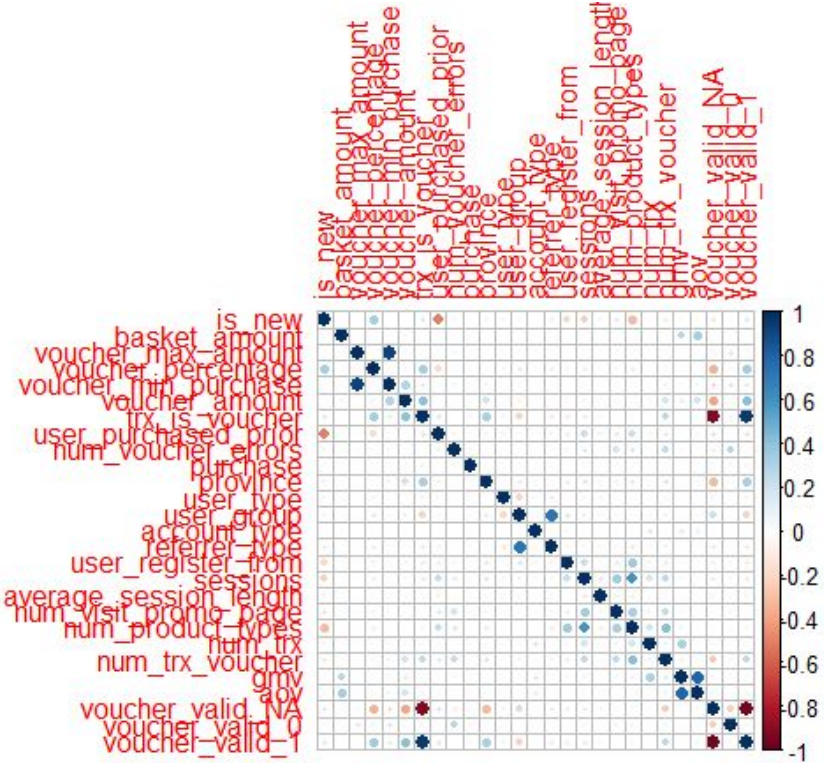
EDA - voucher_valid missing (1)



This is voucher_valid before imputation.

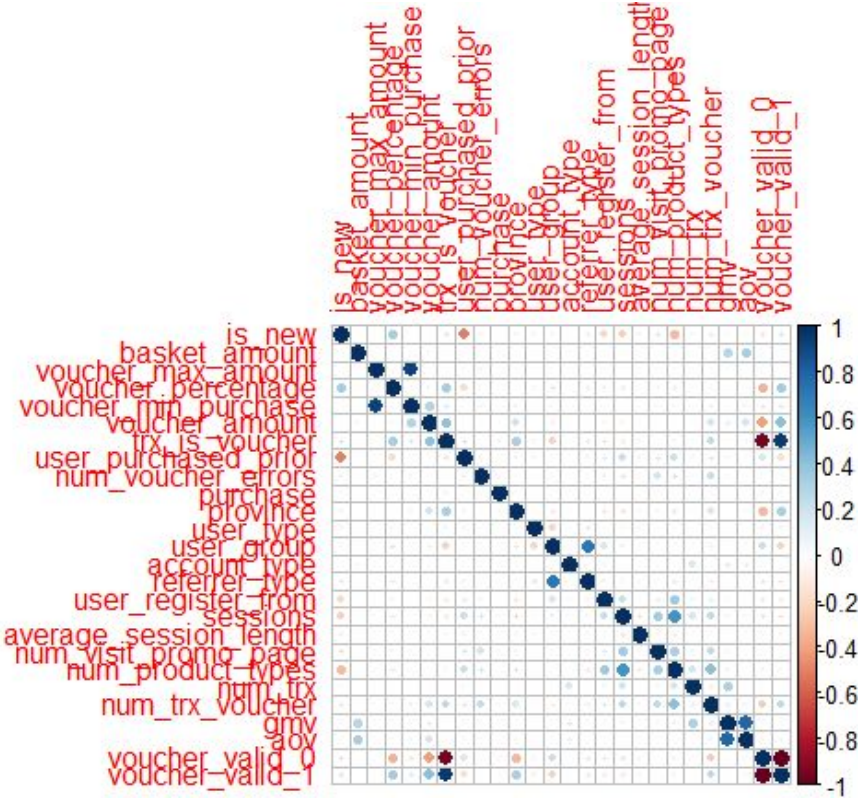
Since voucher_valid is the prequisite of trx_is_voucher, we will impute voucher_valid from trx_is_voucher

EDA - voucher_valid missing (2)



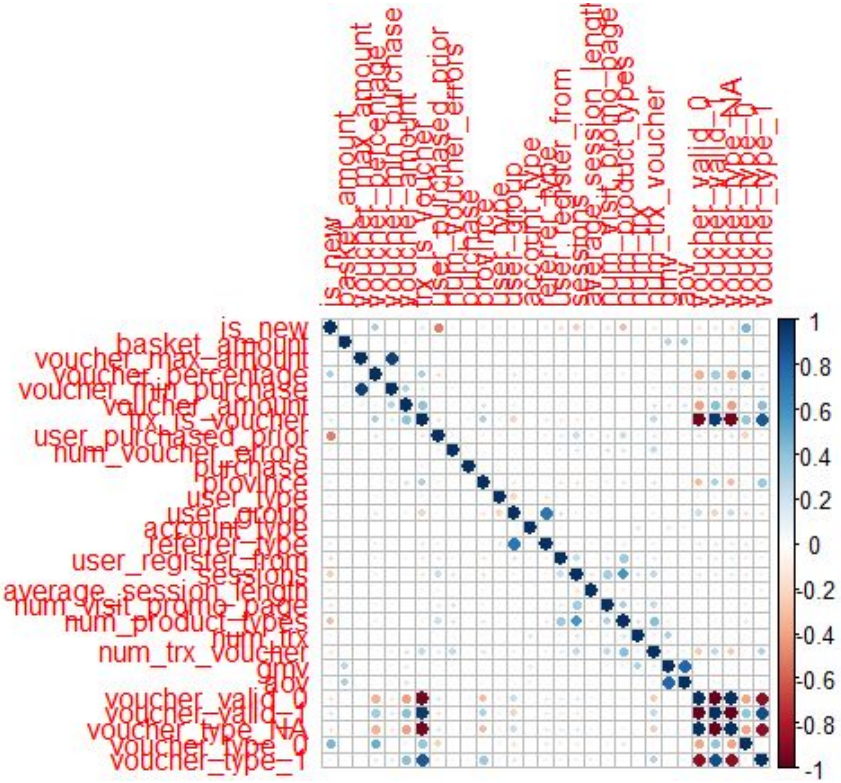
we see that even after imputation, voucher_type_NA is still correlated negatively with trx_is_voucher, so I will impute voucher_type==NA values with 0

EDA - voucher_valid missing (3)



This is voucher_valid after we imputed NA with 0 because of the previous correlation plot

EDA - voucher_type missing



Since voucher_type 0 and 1 positively correlated with purchase, while NA correlated negatively, we consider voucher_type==NA as a third different category that are not fit to type_0 and type_1.

Research Questions

- What are at least 2 factors contribute to users making a purchase?
  - How are the calculation of their potential impact?
- What actions we can take to improve conversion?
- How much impact removing vouchers would have on the business?
  - on basket amount
  - on unique buyers

Variables used for Regression Analysis

Here are the name of variables and their data type:

```
  Dependent Variable (Label)

   $ purchase              : int  0,1

  Independent Variable (Features)

   $ is_new                : int  0,1
   $ voucher_type          : Factor  ["NA", 0, 1]
   $ voucher_valid         : int  0,1
   $ basket_amount         : num
   $ voucher_max_amount    : num
   $ voucher_percentage    : num
   $ voucher_min_purchase  : num
   $ voucher_amount        : num
   $ trx_is_voucher        : int  0,1
   $ user_purchased_prior  : int  0,1
   $ num_voucher_errors    : num
   $ province              : Factor
   $ marketing_tier        : Factor
   $ user_type             : Factor
   $ user_group            : Factor
   $ account_type          : Factor
   $ referrer_type         : Factor
   $ user_register_from    : Factor
   $ sessions              : num
   $ average_session_length: num
   $ num_visit_promo_page  : num
   $ num_product_types     : num
   $ num_trx               : num
   $ num_trx_voucher       : num
   $ gmv                   : num
   $ aov                   : num
   $ ageday: num (time - account_created_at, converted to days)
```

Variables are not used for Regression Analysis

Can only happened after purchase
- is_paid
- Is_remitted

Can be used but not directly:
- account_created_at
- time

Which factors contribute to users making a purchase? (1)

By logistic regression analysis with p-value < 0.001, we check the coefficient. **+** means positive impact, **-** means negative impact

```
> dfmod[dfmod$`Pr(>|z|)` < 0.001 ,]
                      Estimate   Std. Error    z value     Pr(>|z|)
num_trx             433.493501979 40.482098421  10.708276 9.307851e-27
gmv                  61.997439163 16.182056624   3.831246 1.274960e-04
voucher_min_purchase -20.580556025  4.871066152  -4.225062 2.388751e-05
voucher_max_amount    20.073916622  4.884510210   4.109709 3.961577e-05
sessions              -4.336278947  0.312526134 -13.874932 8.987791e-44
province17             2.791850167  0.508848053   5.486609 4.097233e-08
province16             2.666802293  0.716056485   3.724290 1.958654e-04
province15            -2.330526998  0.313871461  -7.425100 1.126947e-13
user_group1            2.281896090  0.643445217   3.546372 3.905745e-04
referrer_type1        -2.200019839  0.643917681  -3.416617 6.340450e-04
referrer_type2        -2.199120750  0.643286858  -3.418569 6.295127e-04
num_visit_promo_page  -2.132787674  0.310210612  -6.875289 6.186422e-12
marketing_tiertier_3   1.070734281  0.189135161   5.661212 1.503074e-08
province12             1.053367257  0.258068206   4.081740 4.469983e-05
num_product_types      1.012208877  0.157651162   6.420561 1.357732e-10
province13             0.975569183  0.178044159   5.479366 4.268516e-08
user_type9            -0.950447102  0.190085006  -5.000116 5.729580e-07
province14             0.942479647  0.220108022   4.281896 1.853076e-05
voucher_type1          0.926475980  0.144839747   6.396559 1.589175e-10
marketing_tiertier_2   0.887369389  0.172224285   5.152406 2.571659e-07
marketing_tiertier_1   0.839380213  0.173356242   4.841938 1.285788e-06
is_new                 0.737891074  0.054811457  13.462351 2.604941e-41
province2              0.658489856  0.057951163  11.362841 6.402791e-30
user_type2            -0.608995485  0.062739191  -9.706779 2.821103e-22
province4              0.520217099  0.058616070   8.874991 6.993904e-19
province8              0.516096449  0.105673989   4.883855 1.040315e-06
user_purchased_prior   0.402888699  0.027676072  14.557294 5.249528e-48
user_type3            -0.361590123  0.065986519  -5.479757 4.259092e-08
province10             0.361105059  0.107593875   3.356186 7.902539e-04
user_register_from2   -0.300384874  0.067236806  -4.467566 7.911458e-06
user_type4            -0.288506173  0.082716678  -3.487884 4.868594e-04
province3              0.247055379  0.048064946   5.140032 2.746911e-07
user_type1            -0.218017553  0.041581861  -5.243093 1.579071e-07
province1              0.215363180  0.038939451   5.530719 3.189200e-08
account_day            0.105026655  0.017701299   5.933274 2.969529e-09
num_voucher_errors    -0.035436133  0.002562339 -13.829602 1.689529e-43
voucher_percentage    -0.009857412  0.001731419  -5.693257 1.246387e-08
```

Brief Description:

Num_trx (+): users who have transacted are strongly more likely to purchase
Gmv (+): users with more gmv are strongly more likely to purchase
Voucher_min_purchase -: higher voucher_min_purchase reduce purchase
Voucher_max_amount (+): lower voucher_max_amount increase purchase
Sessions (-): more sessions does not increase purchase
Users in these **Province:**

- **(+)** More likely to purchase: 17, 16,12,14,13,3,4,8,2,10,3,1
- **(-)** Less likely to purchase: 15

Users in these **User_group**:

- **(+)** More likely to purchase: 1:

Users with these **referrer**:

- **(-)** Less likely to purchase: 1,2

Num_visit_promo_page (-):does not increase purchase
Users in **Marketing_tier (+)** [1,2,3] are more likely to purchase
Num_product_types (+) increase the likelihood to purchase
Users with these **User_type:**

- **(-)** Less likely to purchase: 9,2,3,4,1

Users with **Voucher_type1 (+)** are more likely to purchase
Is_new (+): new users are more likely to purchase
User_purchased_prior (+) in the month, are more likely to purchase
User_register_from2 (-) are less likely to purchase than the other source
Ageday (+): more older the account, more likely to purchase
num_voucher_errors (-): correlate negatively to purchase
Voucher_percentage (-): does not increase purchase

Which factors contribute to users making a purchase? (2)

The two main factors are num_trx & gmv, which are the sign of a mature user. But num_trx & gmv are historical features, and cannot be treated real time.

So, the two main factors which could be treated in real time are:
voucher_min_purchase & voucher_max_amount

I will explain the calculation of impact in the next slide

Calculation of their potential impact

	Estimate	exp(coef)	exp(coef)-1
num_trx	433.493502	1.835844e+188	1.835844e+188
gmv	61.997439	8.416775e+26	8.416775e+26
voucher_min_purchase	-20.580556	1.153395e-09	-1.000000e+00
voucher_max_amount	20.073917	5.223856e+08	5.223856e+08

Increase of 10% in voucher_min_purchase, associated with decrease of 10% in purchase

```
> sort(unique(df$voucher_min_purchase))
 [1] 0.00000 0.00001 0.00002 0.00040 0.00100 0.00120 0.00200 0.00300
 [9] 0.00400 0.00600 0.01000 0.02000 0.04000 0.10000 0.20000 1.00000
>
> sort(unique(df$voucher_max_amount))
 [1] 0.0e+00 2.0e-07 1.0e-06 2.0e-06 3.0e-06 4.0e-06 5.0e-06 6.0e-06
 [9] 1.0e-05 1.1e-05 1.4e-05 2.0e-05 3.0e-05 4.0e-05 5.0e-05 6.0e-05
[17] 8.0e-05 1.0e-04 2.0e-04 3.0e-04 4.0e-04 5.0e-04 1.0e-03 2.0e-03
[25] 3.0e-03 1.0e+00
```

Increase of 0.000001 in voucher_max_amount, associated with increase of 522.3856x in purchase

What actions we can take to improve conversion?

- Retain Loyal users (users with high num_trx, gmv, account_age_day) because they are strongly more likely to make purchase
- Get users to buy different type of products
- Get users to at least do one purchase in a month, because they are likely to purchase again. Try to do push notification
- On vouchers:
    - Reduce Voucher_min_purchase
    - Increase Voucher_max_amount
    - Boost voucher type=1
    - Less priority on voucher_percentage, instead, focus on the things above
- Actively attract new users within this group, because they are more likely to purchase:
    - Province [17, 16,12,14,13,3,4,8,2,10,3,1]
    - User_group [1]
    - Marketing_tier [1,2,3]
- Less priority on attracting users from these group for now, and find out why they are not likely to purchase:
    - Province 15
    - User_type (9,2,3,4,1)
    - User_register_from 2
- Decrease num_voucher_errors, maybe by UI improvement
- Evaluate the referrer 1&2 which has negative correlation
- No need to make users to visit promo page repeatedly
- Find out why the number of sessions correlates negatively. But might be because users are just looking.

Impact of removing vouchers: Voucher Rate Analysis

To measure the impact of voucher for each user, I developed a metric named voucher_rate which is:

`num of purchase_with_voucher / num of purchase`

then bin the users into 10 deciles

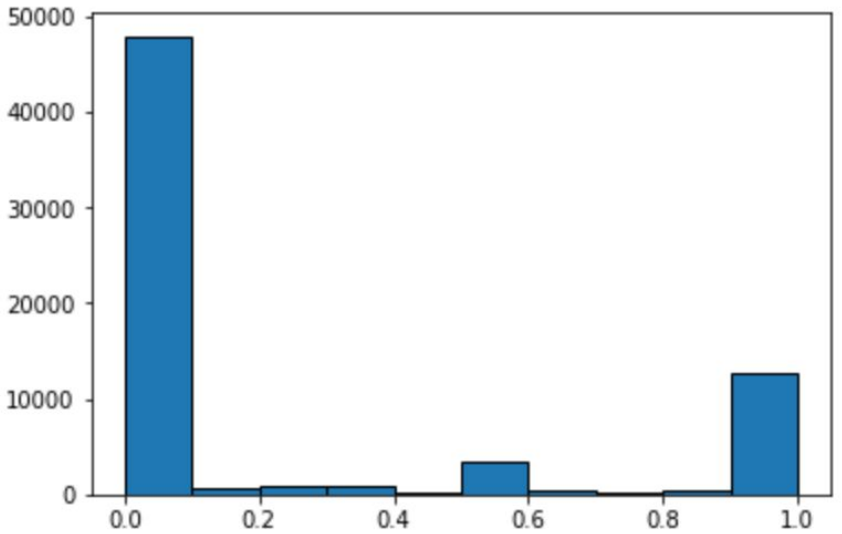
```
cpurchased = df[df.purchase==1]
cvoucher = cpurchased[cpurchased.trx_is_voucher==1].groupby('user_id').purchase.sum()
call = cpurchased.groupby('user_id').purchase.sum()
```

```
voucher_rate=(cvoucher/call).fillna(0).rename('voucher_rate')
```

```
cpurchased=cpurchased.join(voucher_rate.rename('voucher_rate'), on='user_id')
```

```
import matplotlib.pyplot as plt

num,bin,_=plt.hist(voucher_rate, bins=10, edgecolor="k")
```



Impact of removing vouchers on basket amount

bin	withvoucher	nonvoucher	difference	%
(-0.001, 0.1]	0.000067	0.000067	3.326690e-08	0.050000
(0.1, 0.2]	0.000092	0.000089	2.175286e-06	2.373295
(0.2, 0.3]	0.000105	0.000094	1.102453e-05	10.487412
(0.3, 0.4]	0.000118	0.000082	3.609190e-05	30.636798
(0.4, 0.5]	0.000085	0.000065	1.999175e-05	23.623417
(0.5, 0.6]	0.000174	0.000112	6.216741e-05	35.701060
(0.6, 0.7]	0.000277	0.000106	1.710720e-04	61.834185
(0.7, 0.8]	0.000280	0.000135	1.450545e-04	51.742123
(0.8, 0.9]	0.000217	0.000105	1.121365e-04	51.608604
(0.9, 1.0]	0.000095	0.000057	3.718153e-05	39.308087

Users voucher_rate are binned into 10 deciles, these are the basket_amounts we are likely to lose when we remove vouchers.

The drop in basket_amount is increasing with the users voucher rate until the (0.6,0.7] bin

Impact of removing vouchers on unique users

bin	withvoucher	nonvoucher	difference	%
(-0.001, 0.1]	48053	48053	0	0.000000
(0.1, 0.2]	786	786	0	0.000000
(0.2, 0.3]	621	621	0	0.000000
(0.3, 0.4]	1051	1051	0	0.000000
(0.4, 0.5]	3423	3423	0	0.000000
(0.5, 0.6]	165	165	0	0.000000
(0.6, 0.7]	495	495	0	0.000000
(0.7, 0.8]	395	395	0	0.000000
(0.8, 0.9]	310	310	0	0.000000
(0.9, 1.0]	12618	122	12496	99.033127

Users voucher_rate are binned into 10 deciles, these are the unique users we are likely to lose when we remove vouchers.

We are very likely to lose 12496 users who uses voucher more than 90% on their transaction, while the rest of users have no difference

Thank you