Objective

The NextGen shopping giant has divided its operation in Singapore into 2 zones (East/West). Through their pos systems they consolidate customer purchase data to keep an eye on what customers like to buy for daily usages and what they buy occasionally. Their data science team is looking to build recommendations to facilitate their customers' future purchases by identifying sets of items that can be offered as bundles. The objective of this project is to create customer specific approaches of cross and up selling.

Contents

1 Introduction

In this project, we will apply market basket analysis to analyse NetGen shopping giant's customer purchase data in pos systems and combine the analysis result with customer profile to create customer specific approaches of cross selling and up selling.

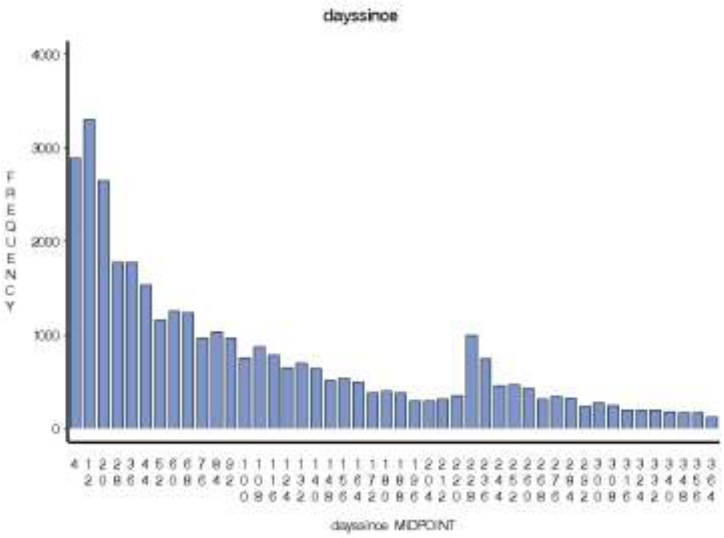
2 Data Preparation

2.1 Data clean

There are 12 rows where the column "dayssince" is negative value "-2576980378". It should be a system errors, therefore we removed these 12 rows from dataset.

Furthermore, after removing the negative values, we noticed that the range of column "dayssince" is from 2 to 437 days. Since using out of date data may affect our adjudgment and outcome accuracy, we only focus on past one year's data and removed the rows which "dayssince" is greater than 365 days.

The distribution of "dayssince" after data clean is shown as below.



2.2 Data transformation

Since we are going to apply market basket analysis to this dataset, we need to transform it into transaction data.

We create a new copy of cleaned dataset and stack the 7 columns named "Hill Climb Racing", "Batman: Arkham Knight (PS4)", "The Witcher 3: Wild Hunt (PS4)", "Graphic Card", "Subway Surfer", "Sound Booster", "Wireless Controller" to one column and name it as "item". The value of each item is captured in column named "quantity". Then we deleted the rows which quantity is equal to zero.

The example is shown in the table below.

Original data:

CID	Zone	Sex	Age	Days Since	Dollar Amount	Trans Count	Hill Climb Racing	Batman: Arkham Knight (PS4)	The Witcher 3: Wild Hunt (PS4)	Graphic Card	Subway Surfer	Sound Booster	Wireless Controller
C1040	East-Zone	0	36	72	268	5	2	0	0	0	1	1	0
C1041	West-Zone	1	34	3	993	27	16	0	2	0	5	0	0

After transformation:

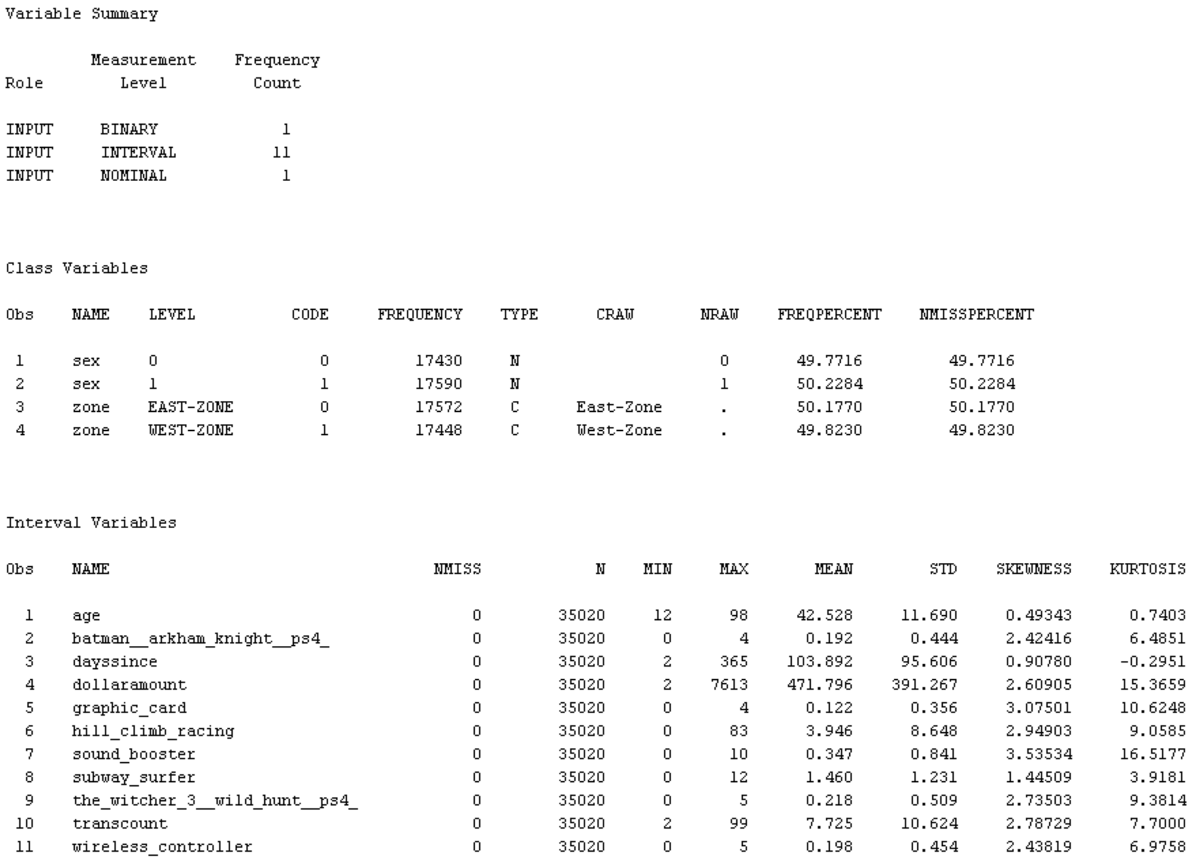
CID	Zone	Sex	Age	Days Since	Dollar Amount	Trans Count	item	quantity
C1040	East-Zone	0	36	72	268	5	Hill Climb Racing	2
C1040	East-Zone	0	36	72	268	5	Subway Surfer	1
C1040	East-Zone	0	36	72	268	5	Sound Booster	1
C1041	West-Zone	1	34	3	993	27	Hill Climb Racing	16
C1043	West-Zone	1	34	3	993	27	The Witcher 3: Wild Hunt (PS4)	2
C1045	West-Zone	1	34	3	993	27	Subway Surfer	5

Since we only focus on the transaction information, we removed the unnecessary columns and only keep the following 3 columns.

Column	Role
CID	ID
item	Target
quantity	Input

2.3 Exploratory data analysis

The summary shown below are the variables and their distribution of original data after data clean and transformation. Each customer has 7. 7 transactions in last year on average. The small game hill climb racing is the best seller and has average purchase times as high as 3.946 times, then followed by small game subway surfer whose average purchase time is 1.46.



```
Variable Summary

         Measurement    Frequency
Role        Level         Count

INPUT     BINARY            1
INPUT     INTERVAL         11
INPUT     NOMINAL           1


Class Variables

Obs    NAME    LEVEL          CODE    FREQUENCY    TYPE      CRAW       NRAW    FREQPERCENT    NMISSPERCENT

 1     sex     0                 0        17430     N                    0        49.7716         49.7716
 2     sex     1                 1        17590     N                    1        50.2284         50.2284
 3     zone    EAST-ZONE         0        17572     C      East-Zone     .        50.1770         50.1770
 4     zone    WEST-ZONE         1        17448     C      West-Zone     .        49.8230         49.8230


Interval Variables

Obs    NAME                                 NMISS          N     MIN     MAX      MEAN        STD    SKEWNESS    KURTOSIS

  1    age                                      0      35020      12      98     42.528     11.690     0.49343      0.7403
  2    batman__arkham_knight__ps4_              0      35020       0       4      0.192      0.444     2.42416      6.4851
  3    dayssince                                0      35020       2     365    103.892     95.606     0.90780     -0.2951
  4    dollaramount                             0      35020       2    7613    471.796    391.267     2.60905     15.3659
  5    graphic_card                             0      35020       0       4      0.122      0.356     3.07501     10.6248
  6    hill_climb_racing                        0      35020       0      83      3.946      8.648     2.94903      9.0585
  7    sound_booster                            0      35020       0      10      0.347      0.841     3.53534     16.5177
  8    subway_surfer                            0      35020       0      12      1.460      1.231     1.44509      3.9181
  9    the_witcher_3__wild_hunt__ps4_           0      35020       0       5      0.218      0.509     2.73503      9.3814
 10    transcount                               0      35020       2      99      7.725     10.624     2.78729      7.7000
 11    wireless_controller                      0      35020       0       5      0.198      0.454     2.43819      6.9758
```

Firstly, let’s look at product and how often did customer buy each product.

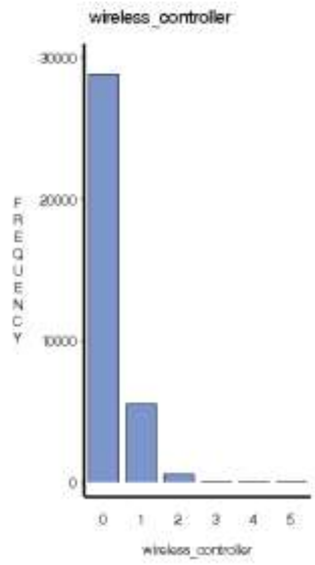
There are 3 product categories and 7 products which the merchant sells. 3 hardware, wireless controller, sound booster and graphic card, 2 PS4 games, Witcher 3: Wild Hunt (PS4) and Batman: Arkham Knight (PS4), 2 small games’ top-up, subway surfer and hill climb racing. We group the products into their product category groups.

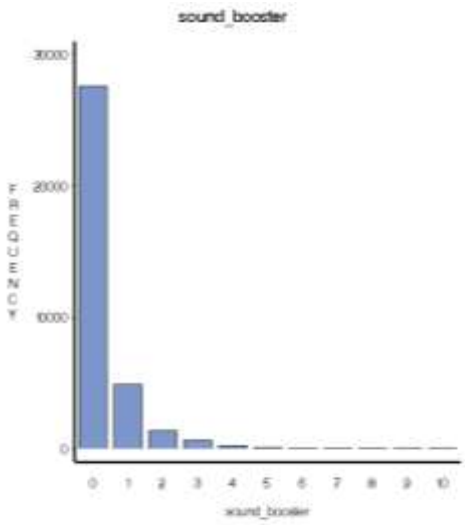
Combine with product nature and distribution, we noted that 42% of customers have bought PC hardware before and very seldom of people bought more than once as normally people only own one PC and people may buy the PC hardware only when their parts are faulty or want to upgrade. Similarly, 32% of customers have bought PS4 games as the games have dependency on hardware and only small amount of people own PS4 and normally people only bought the game once. Therefore, PC hardware and PS4 games are the items customer occasionally buy.

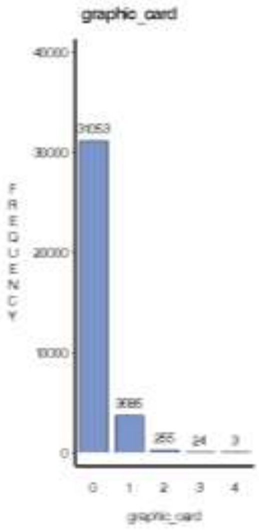
In contrary, 92% of customer has bought the small gamres’ top-up before, and maximum purchase frequency in past year is as high as 87 times. This is beacsue top up the coins continiously in games to buy equipment is a common thing to the players.

As a result, with the differnet nature of the products, the purchasing frequencey varies a lot.

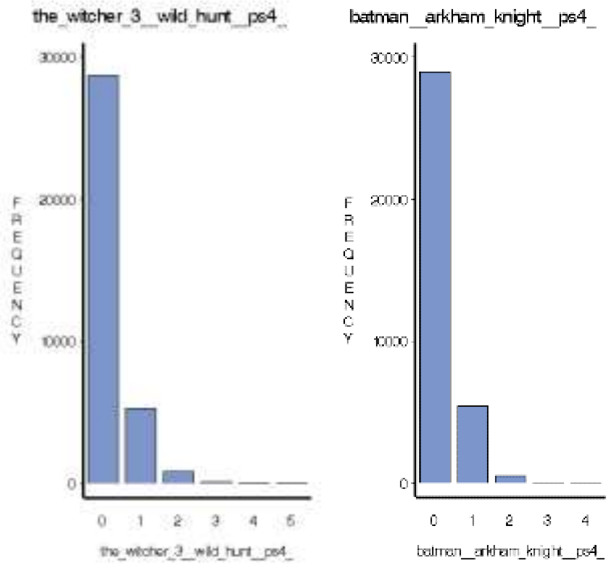
Distribution of hardware



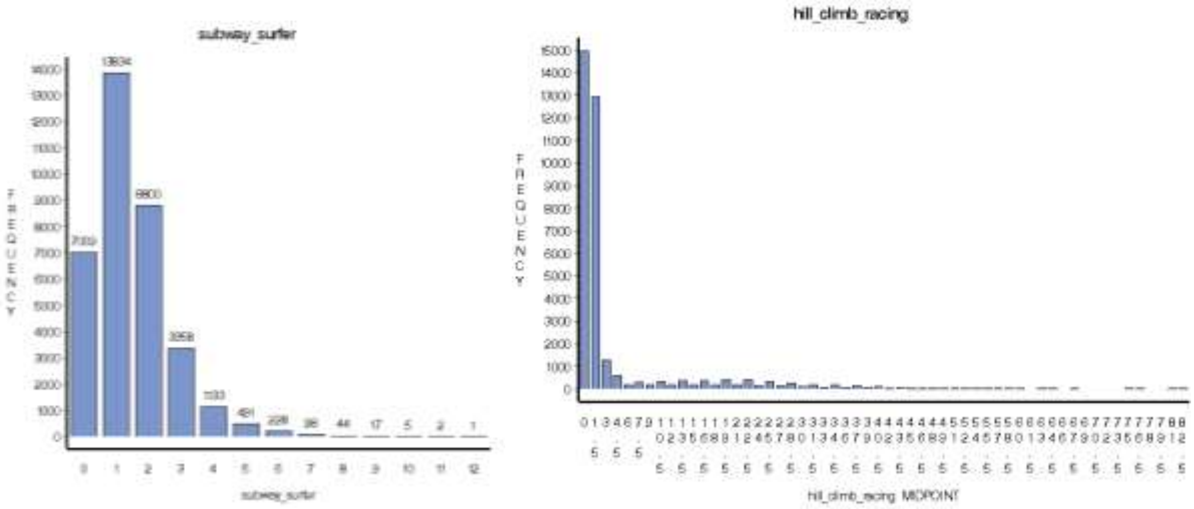




Distribution of PS4 games



Ditribution of small games' top-up



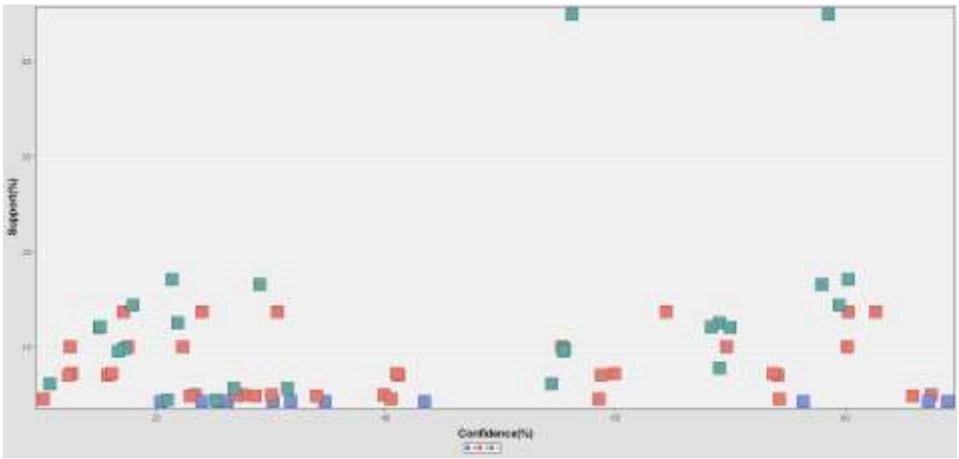
Secondly, let’s look at the whether the zone, sex and age affect the number of purchase times of each products. We made a two sampe T test of each product with both gender and both zones to see if there is a signficant different between different gender and different zones.

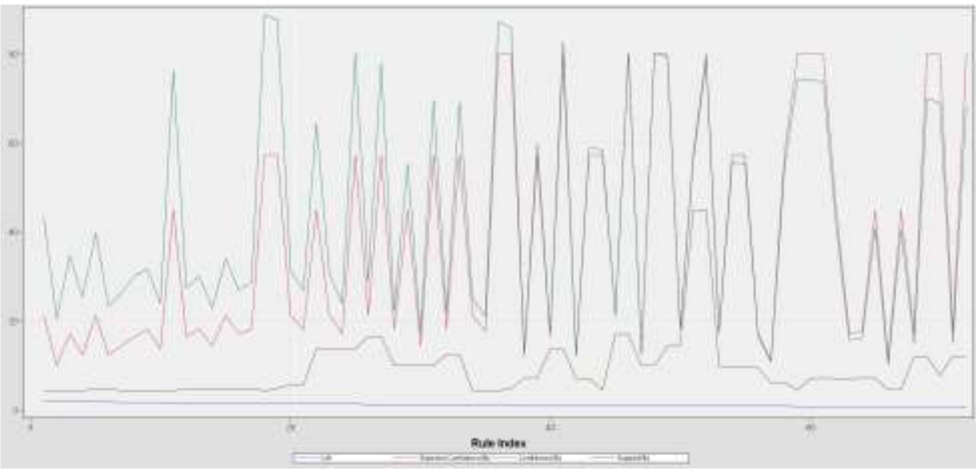
As a result, we found that there is no significant difference between male customers and female customers on their purchasing preference. Similarly, there is also no significant difference between the customers in west zone and east zone on their purchasing preference.

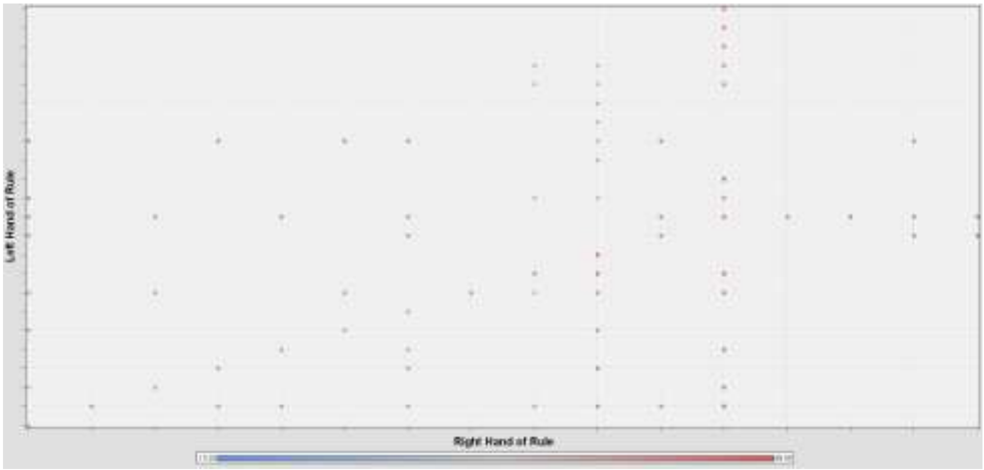
3 Market basket analysis

We applied market basket analysis for the transaction data. The maximum items in each rule was set to 2 and minimum confidence level of each rule was set to 10%. Last but not least, we only look at the rules whose value is greater than 1 as we only focus on the positive relationship between left side and right side.

The statistics plot, statistics line plot and rule matrix are shown below.







From the information above, we found that the rules which have high confidence and support are hill climb racing and subway surfer, which are small games top ups and customer bought in a daily basis. Furthermore, they also appear quite often on the right-hand side. Since they are meaningless for us, we ignore all rules with small games top up hill climb racing and subway surfer on the right-hand side.

Next, we conduct a Chi-Square test for each rule to test the significant. The rules whose p value is greater than 0.0005 will be remove from the list.

The remaining list of rules are shown in the table below.

EXP_CONF	CONF	SUPPORT	LIFT	RULE	index	CHISQ	PVALUE
21.34	39.86	5	1.87	the_witcher_3__wild_hunt__ps4_ & hill_climb_racing ==> sound_booster	5	1026.779	0
18.16	31.72	4.36	1.75	subway_surfer & sound_booster & hill_climb_racing ==> the_witcher_3__wild_hunt__ps4_	9	689.9902	0
18.16	30.07	5	1.66	sound_booster & hill_climb_racing ==> the_witcher_3__wild_hunt__ps4_	13	665.659	0
21.34	33.99	4.91	1.59	the_witcher_3__wild_hunt__ps4_ & subway_surfer ==> sound_booster	15	563.7091	0
18.16	28.64	4.91	1.58	subway_surfer & sound_booster ==> the_witcher_3__wild_hunt__ps4_	17	534.6612	0
21.34	31.46	5.71	1.47	the_witcher_3__wild_hunt__ps4_ ==> sound_booster	20	474.3812	0
18.16	26.78	5.71	1.47	sound_booster ==> the_witcher_3__wild_hunt__ps4_	21	474.3812	0
21.34	25.25	4.48	1.18	wireless_controller ==> sound_booster	34	68.96781	1.11E-16
17.76	21.01	4.48	1.18	sound_booster ==> wireless_controller	35	68.96781	1.11E-16

From the above table, we find that beside the small games, sound booster, PS4 game the witcher 3 wild hunt and wireless controller also appeared in the list of rules. For sound booster, it appears very often with PS4 game the wither 3 wild hunt. Wireless controller and sound booster implied each other and appeared together.

We look at the rule **the_witcher_3__wild_hunt__ps4_ ==> sound_booster** and **sound_booster ==> the_witcher_3__wild_hunt__ps4_**, its support is 5.71%, and confidence is 31.46% and 26.78% which means 5.71% of customer have bought hill climb racing and sound booster together and there is 31.46% and 26.78% of the chance when we see the witcher 3 wild hunt ps4, we also see sound booster or when we see sound booster, we also see the witcher 3 wild hunt ps4.

And we also look at the rule **wireless_controller ==> sound_booster** and **sound_booster ==> wireless_controller**, its support is 4.48 %, and confidence is 25.25% and 21.01% which means 4.48% of customer have bought wireless controller and sound booster together and there is 25.25% and 21.01% of the chance when we see wireless controller, we also see the sound booster or when we see sound booster we also see wireless controller.

4 Result and recommendation

Base on the market basket analysis and customer's need, there are some recommendations on cross selling and up selling.

Since sound booster and the witcher 3 wild hunt (PS4) often go together, put an expensive alternative to the witcher 3 wild hunt (PS4) which is Batman: Arkham Knight (PS4), near the display for sound booster can create an up-sell opportunity. Alternatively, we may put them on sale at different times to drive purchases continually.

Similarly, since sound booster and wireless controller often go together, put a more expensive sound booster near the display for wireless controller can create an up-sell opportunity. Alternatively, we may put them on sale at different times to drive purchases continually.

Last but not least, base on product category, since there are not many customers own PS4, to attract the limit number of customers who have PS4 to buy PS4 games, we can make the two PS4 games into a bundle and give discount to them to create cross sell opportunity. Similarly, since subway surfer and hill climb racing are in the same product category, customer who bought them before have interest in playing small games, we when customer buy one small game, we should also recommend another small game to create cross sell opportunity.