Assignment 3

Q1 Hero	sum_price_rev1	sum_price_rev2	sum_price_rev3	purch_amount_total	purch_days_no_total
Whale (n= 171) (1.4%)	Mean: 64853.801	Mean: 6344.324	Mean: 78923.977	Mean: 1671485.38	Mean: 22.67251
	n=171 (1.4%)	n= 4713	n=171	S.D :856254.9	S.D: 11.24007
Dolphin (n=1933)2104 (16.2%)	Mean: 38760.476	Mean: 49235.359	Mean: 48220.826	Mean: 294427.83	Mean: 9.150665
	n=1933 (16.2%)	n =1827	n= 1671	S.D: 205750.6	S.D: 7.679977
Minnow (n=9796) (82.4%)	Mean: 6344.324	Mean: 9450.032	Mean: 8550.133	Mean: 15810.74	Mean: 2.13383
	n= 9796 (82.4%)	n= 4713	n=2623	S.D: 20824.9	S.D: 1.75659
Total (n=11900)	Mean: 12450.67	Mean:22054.69	Mean: 26091.83	Mean: 84860.08	Mean: 3.37
	n= 11900	n= 6711	n= 4465	S.D: 254837.6	S.D: 4.49
				p1: 1181010	
				p10: 203000	
				p50: 10000	
There are 1.4% of whale users, 16.2% of Dolphin users and 82.4% of minnow users.				Whale users are spent the most money (mean= 1671485.38 KRW) on hero game. While Minnow users are spent the least money (mean= 15810.74 KRW) on hero.	Whale users are spent the most time (mean=22 days) on purchase hero game. While Minnow users are spent the least (mean =2 days) on hero.

Comparison to the groups

	count	mean	sd	sum expense	Findings
whale	171 (1.4%)	1671485.38	856254.9	285824000 (28%)	There are only 1.4% of whale users and they contributed 28% to the expense.
others	11729 (98.6)	61728.28	134245.4	724011000 (72%)	
Dolphin	2104 (18%)	406346.48	489712.6	854953000 (85%)	There are 18% of Dolphin users and they contributed 85% to the expense.
others	9796 (82%)	15810.74	20824.9	154882000 (15%)	
Minnow	9796 (82%)	15810.74	20824.9	154882000 (15%)	There are 82% of minnow users and they contributed 15% to the expense.
others	2104 (18%)	406346.48	489712.6	854953000 (85%)	
P1	119 (1%)	1921680.67	920595.7	228680000 (23%)	For P1, they contributed 23% to the expense.
others	11781 (99%)	66306.34	150604.9	781155000 (77%)	
P10	1193 (10%)	610456.83	570996.51	728275000 (72%)	For P10, they contributed 72% to the expense.
others	10707 (90%)	26296.82	40799.56	281560000 (28%)	

To identify high spenders, we can obtain the information from initial usage and purchase patterns.

Q2 Hero Connection)	no_purch a_days	no_purc ha	sum_expe nse	no_conne ction_before _purch	connection _total_before _purch	freq_conne ction_before_ purch	connection mean_before _purch	connection _var_before_ purch	connection_i ncrease_before _purch
Min	1	1	3000	1	0	0.006173	0	0	0.1833
1st Qu	1	1	3000	10	50	0.666667	3.889	5.076	1.0000
Median	2	2	10000	26	153	0.887097	5.946	11.188	1.0667
Mean	3.374	4.391	84860	35.62	275	0.778523	7.006	19.295	1.6579
3rd Qu	4	5	49000	50	358	0.961538	8.91	22.784	1.7500
Max	74	116	6435000	164	3331	1	44.5	991.982	32.0000
sd	4.487	6.951	254837.6	33.03	380	0.24	4.4	29	1.6894
								NA: 1174	NA:22
Findings (Initial usages of onnection)	Average 3.374 days for user purchased items.	Average 4.391 items of purchases.	Average total expenditure is 84860 KRW	Average no of onnection is 35.62	Average total connection before purch is 275	Average frequency of connection before purch is 0.78	Average connection mean before purch is 7	Average onnection ar before urch is 19.3	Average onnection ncrease before urcch is 1.66

Cont' Q2 Hero (Connection)	no_connection _before_purch	connection_tota l_before_purch	time_to_pur ch	freq_connection _before_purch	connection_mean _before_purch	connection_var_ before_purch
no_connection_ before_purch	1	0.8237934	0.63656743	0.30573405	0.10500503	0.00055741
connection_tota l_before_purch	0.823793395	1	0.47416816	0.31592843	0.3935145	0.160255008
time_to_purch	0.636567426	0.4741682	1	-0.37102138	-0.08855156	-0.033241929

freq_connection_before_purch	0.305734055	0.3159284	-0.37102138	1	0.34504068	0.060551083
connection_mean_before_purch	0.105005028	0.3935145	-0.08855156	0.34504068	1	0.55446418
connection_var_before_purch	0.00055741	0.160255	-0.03324193	0.06055108	0.55446418	1

From the correlation plot, we found that strong positive correlations in (i) "connection_total_before_purch vs no_connection_before_purch" (r~0.82) and (ii) "time_to_purch vs no_connection_before_purch (r~0.64)". While slightly weak negative correlation in "freq_connection_before_purch vs time_to_purch (r ~ - 0.37)".

Cont' Q2 Hero	no_usage_before_purch	usage_total_before_purch	freq_usage_before_purch	session_mean_before_purch	session_var_before_purch	usage_increase_before_purch
Min	1.00	1.504e+03	0.005917	1504	1.295e+07	0.003
1st Qu	9.00	3.111e+07	0.600000	2459169	5.226e+12	1
Median	25.00	9.889e+07	0.838235	4444430	1.360e+13	1.003
Mean	34.54	2.196e+08	0.743282	5626567	1.883e+13	5.637
3rd Qu	48.00	2.652e+08	0.941176	8025779	2.596e+13	1.754
Max	163.00	2.560e+09	1.000000	34579265	5.022e+14	3387.528
sd	32.86396	3.26e+08	0.2458083	4273770	1.95e+13	66.35
					NA: 1080	

Cont' Q2 Hero	no_usage_before_purch	usage_total_before_purch	time_to_purch	freq_usage_before_purch	session_mean_before_purch	session_var_before_purch	usage_increase_before_purch
no_usage_before_purch	1.00000000	0.7773085	0.636255442	0.31251400	0.12209031	0.090514777	-0.045124348
usage_total_before_purch	0.77730847	1.0000000	0.444790699	0.31039721	0.45192048	0.275370386	-0.033713199
time_to_purch	0.63625544	0.4447907	1.000000000	-0.34207454	-0.05539171	0.005951864	-0.041452060
freq_usage_before_purch	0.31251400	0.3103972	-0.342074535	1.00000000	0.33435063	0.152944770	-0.037127990
session_mean_before_purch	0.12209031	0.4519205	-0.055391708	0.33435063	1.00000000	0.632189612	-0.051757835
session_var_before_purch	0.09051478	0.2753704	0.005951864	0.15294477	0.63218961	1.000000000	0.003267107
usage_increase_before_purch	-0.04512435	-0.0337132	-0.041452060	-0.03712799	-0.05175784	0.003267107	1.000000000

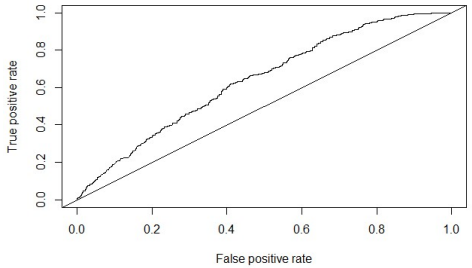
From the correlation plot, we found that strong positive correlations in (i) "usage_total_before_purch vs no_usage_before_purch" (r~0.78) and (ii) "time_to_purch vs no_usage_before_purch" (r~0.64). While slightly weak negative correlation in "freq_usage_before_purch vs time_to_purch" (r ~ - 0.34).

Conclusions: (i) The early time "hero app" usage and spending behaviours of users contained important information to predict the future planning and marketing strategies for three different users groups (whale, dolphin & minnow). (ii). The variables like "purch_amount_total" and "purch_days_no_total" could be highly correlated to high spenders.

Q3 Comparison of classifications models

From Q1, we found that 18% of Dolphin users are contributed 85% to the expense. To correctly identifies the Dolphin users, so we are built several model which are based on usage/connection pattern & sum expense of hero games.

	Description	Advantages	Disadvantages
Recursive partitioning tree	RP tree chooses variables to maximize information gains. Based on entropy measures such as Shannon index and Gini index.	-Very flexible and easy to interpret -Works on both classification and regression problems -Nonparametric	-Prone to bias and overfitting
Conditional inference tree	CI tree adapts the significant test procedures to select variables.	-Very flexible and easy to interpret -Works on both classification and regression problems -Nonparametric - Less prone to bias than a recursive partitioning tree	-Prone to over-fitting
Logistic regression		-Easy to interpret -Provides model logistic probability -Provides confidence interval - quickly update the classification model to incorporate new data	-Suffers multicollinearity -Does not handle the missing value of continuous variables -Sensitive to extreme values of continuous variables

Selected results	Recursive partitioning tree Model B	Conditional Inference Tree Model M1-C	Logit for M1A
Attributes	is_dolphin ~no_connection_before_purch +no_purcha+connection_total_before_purch+ freq_connection_before_purch +connection_mean_before_purch + connection_var_before_purch + connection_increase_before_purch	user_class ~ p1.x + p5.x + p10.x	glm(formula = is_dolphin ~ ., family = "binomial", data = M1A_logit)
Root node error	42369/236497 = 0.17915		AUC 0.8047039
Accuracy	0.9183		Sensitivity 0.6187779 Specificity 0.5911306 Cut off: 0.2084144
p-value	< 2.2e-16 [Acc > NIR] ; < 2.2e-16 Mcnemar's Test P-Value		
Kappa	0.6888		
Sensitivity	0.9247		
Specificity	0.8760		
Pos Pred Value	0.9804		
Neg Pred Value	0.6340	Findings: I will choose Logistic regression, since it is quickly update the classification model to incorporate new data. Since they are moderately value of Sensitivity 0.6187779 and Specificity 0.5911306. We trade-off between prediction accuracy and timely appositeness for identify dolphin users.	
Prevalence	0.8704		
Detection Rate	0.8048		
Detection Prevalence	0.8208		
Balanced Accuracy	0.9003		
Min cross-validation error	0.4557577		
Locate min cross-validation error	2		
cost complexity parameter	0.01		

```
Model formula:
user_class ~ p1.x + p5.x + p10.x

Fitted party:
[1] root
|   [2] p10.x <= 0: minnow (n = 210677, err = 7.9%)
|   [3] p10.x > 0
|   |   [4] p1.x <= 0
|   |   |   [5] p5.x <= 0: dolphin (n = 11287, err = 0.0%)
|   |   |   [6] p5.x > 0: dolphin (n = 11321, err = 7.9%)
|   |   [7] p1.x > 0: whale (n = 3212, err = 0.0%)

Number of inner nodes:    3
Number of terminal nodes: 4
null device
          1
```


Predictions of is_dolphin		
	No	Yes
No	190327	3801
Yes	15509	26860