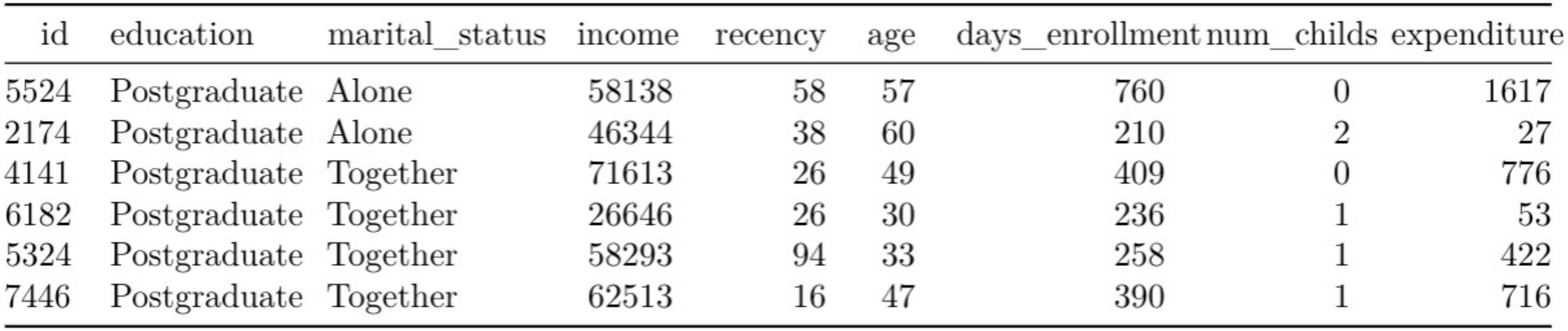
Customer Personality Analysis

Presenter: Xinxin Chen

Date: 11/29/2021

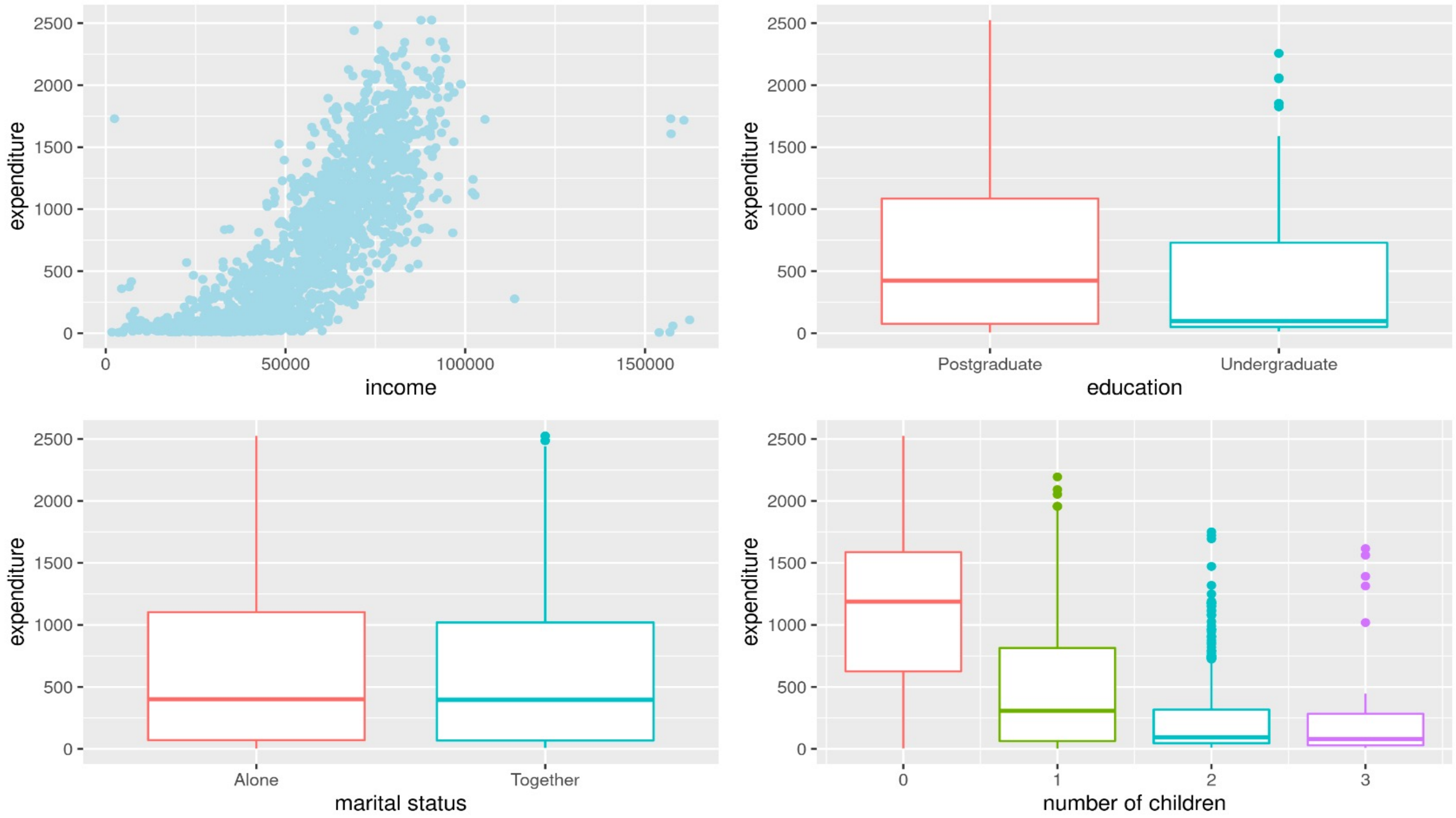
Background



id	education	marital_status	income	recency	age	days_enrollment	num_childs	expenditure
5524	Postgraduate	Alone	58138	58	57	760	0	1617
2174	Postgraduate	Alone	46344	38	60	210	2	27
4141	Postgraduate	Together	71613	26	49	409	0	776
6182	Postgraduate	Together	26646	26	30	236	1	53
5324	Postgraduate	Together	58293	94	33	258	1	422
7446	Postgraduate	Together	62513	16	47	390	1	716

- Data source: https://www.kaggle.com/imakash3011/customer-personality-analysis
- Dimension: 2240 × 29, each row representing the data from a unique customer.
- Variables: income, education level, marital status, the amount of money spent on wines, fruits, and meat during 2012-2014, ...

Initial Analysis

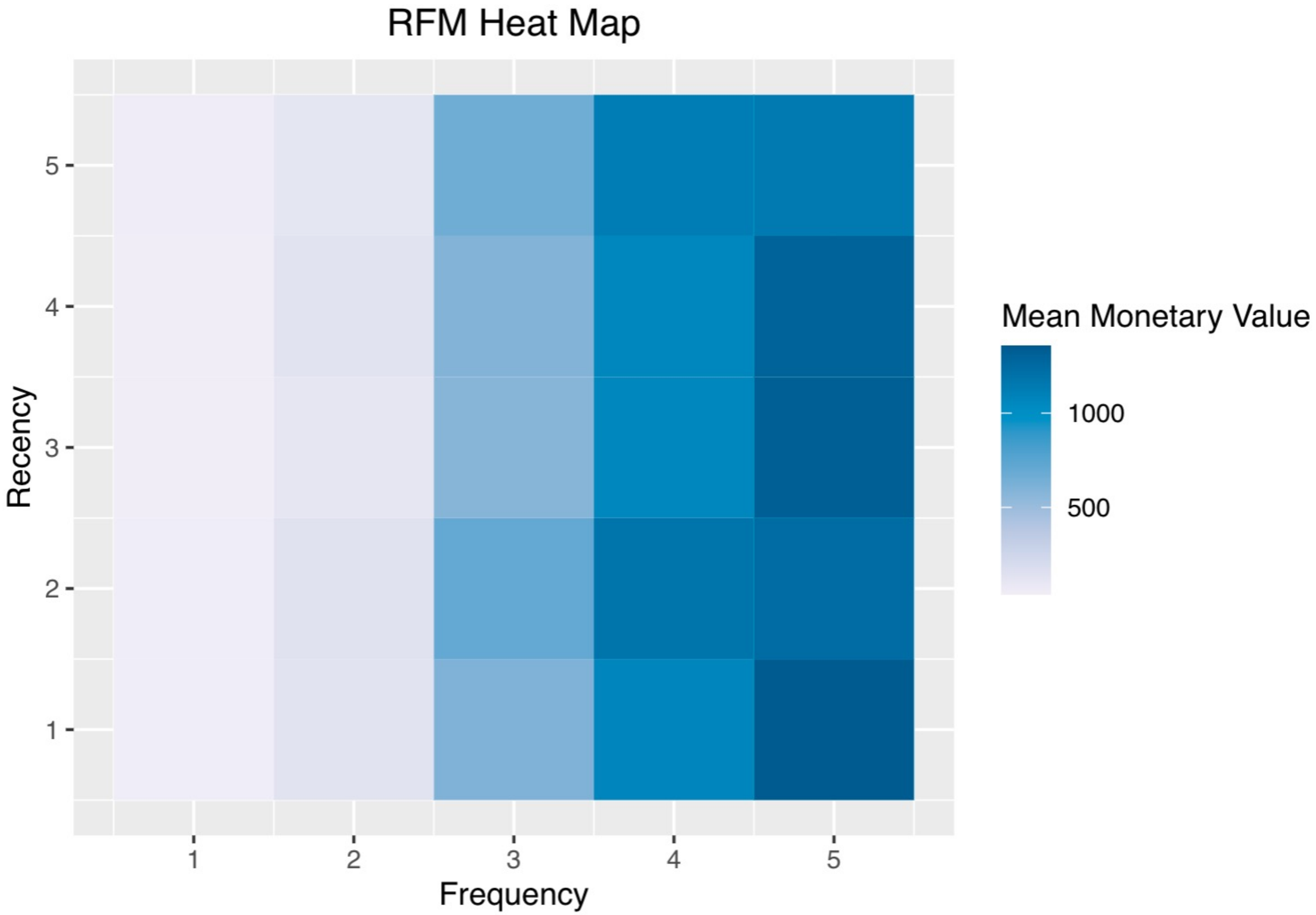


RFM Analysis

- RFM stands for Recency, Frequency, and Monetary and is used to categorize customers based on these three quantitative factors.
- Recency, frequency, and monetary scores (ranged from 1-5) are assigned to each customer. Higher scores usually represent better customers. The ideal customer would have a high score in each factor.
- RFM score is generated by concatenating these three scores into a single value.

RFM Analysis

- Monetary Value is concentrated in the high frequency



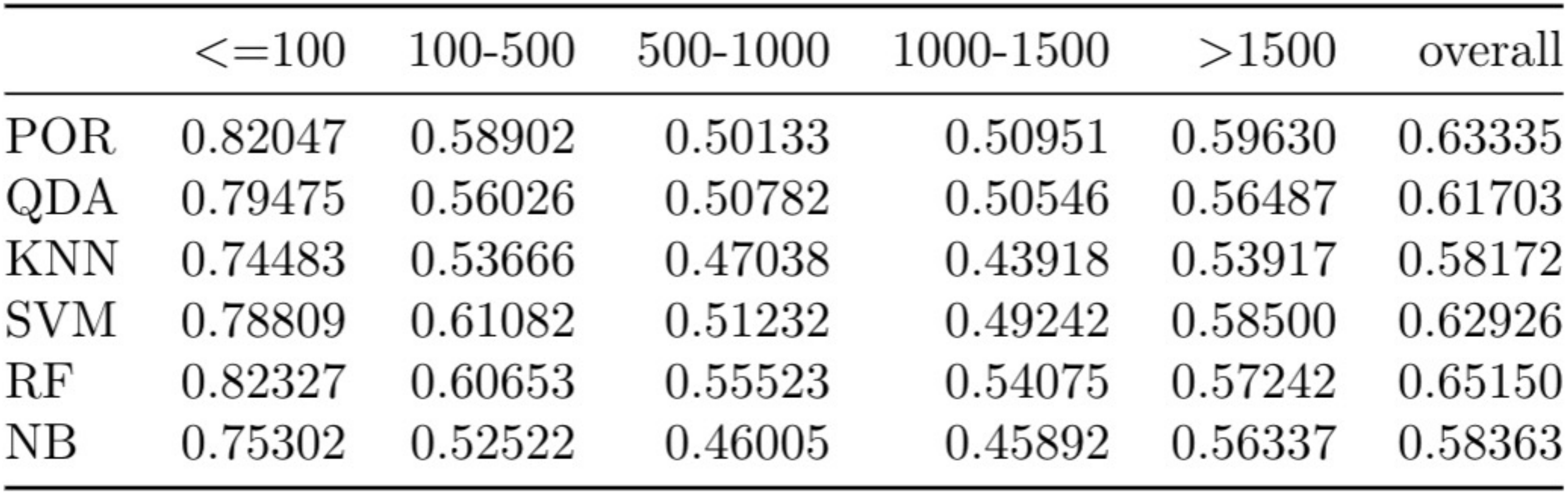
Segment

Segment	Description	R Score	F Score	M Score	Customer Percentage
Loyal Customers	Spend good amount, bought frequently	2-5	3-5	3-5	34.09%
Potential Loyalist	Recent customers, spent good amount, bought more than once	3-5	1-3	1-3	25.41%
Need Attention	Average recency, frequency & monetary values	2-3	2-3	2-3	3.07%
At Risk	Spent good amount, purchased often but long time ago	<=2	2-5	2-5	14.69%
Lost	Lowest recency, frequency & monetary scores	<=2	<=2	<=2	5.11%
	……				

Prediction of Expenditure

- Convert the expenditure variable into a nominal variable with 5 levels: <=100, 100-500, 500-1000, 1000-1500, >1500.
- Predict expenditure based on customers’ education level, marital status, income, age, days since enrollment with the company, and number of children.
- Models considered: proportional odds model (POR), quadratic discriminant analysis (QDA), k-nearest neighbors (KNN) with k = 15, support vector machine (SVM), random forests (RF), and naive bayes (NB).

Model Prediction Accuracy Comparison



	<=100	100-500	500-1000	1000-1500	>1500	overall
POR	0.82047	0.58902	0.50133	0.50951	0.59630	0.63335
QDA	0.79475	0.56026	0.50782	0.50546	0.56487	0.61703
KNN	0.74483	0.53666	0.47038	0.43918	0.53917	0.58172
SVM	0.78809	0.61082	0.51232	0.49242	0.58500	0.62926
RF	0.82327	0.60653	0.55523	0.54075	0.57242	0.65150
NB	0.75302	0.52522	0.46005	0.45892	0.56337	0.58363