What done in the original solution

Data preparing

1. Quantity > 0
2. 'CustomerID' not null
3. 'InvoiceDate' < '1/12/2011'
4. Create a 'Sales' Variable = 'Quantity' × 'UnitPrice'
5. Summarize the 'Sales count' (frequency), 'Recency', and 'total amount', 'avg. number of days between purchases'

Basic Data information

1. Data Range: 1/12/2010 - 9/12/2011 (original)
2. Data Range: 1/12/2010 – 1/12/2011 (new)
3. There are 1539 will have only 'Sale count'= 1, and filter them out
4. 'orders_df'

		Sales	InvoiceDate
CustomerID	InvoiceNo		
12346.0	541431	77183.60	2011-01-18 10:01:00
12347.0	537626	711.79	2010-12-07 14:57:00
	542237	475.39	2011-01-26 14:30:00
	549222	636.25	2011-04-07 10:43:00
	556201	382.52	2011-06-09 13:01:00
	562032	584.91	2011-08-02 08:48:00

5. 'summary_df'
6.

CustomerID	sales_min	sales_max	sales_sum	sales_avg	sales_count	invoicedate_min	invoicedate_max	invoicedate_purchase_duration	invoicedate_purchase_frequency
12347.0	382.52	1294.32	4085.18	680.863333	6	2010-12-07 14:57:00	2011-10-31 12:25:00	327	54.500000
12348.0	227.44	892.80	1797.24	449.310000	4	2010-12-16 19:09:00	2011-09-25 13:13:00	282	70.500000
12352.0	120.33	840.30	2506.04	313.255000	8	2011-02-16 12:33:00	2011-11-03 14:37:00	260	32.500000
12356.0	58.35	2271.62	2811.43	937.143333	3	2011-01-18 09:50:00	2011-11-17 08:40:00	302	100.666667
12359.0	547.50	2876.85	6372.58	1593.145000	4	2011-01-12 12:43:00	2011-10-13 12:47:00	274	68.500000
...	...	...	...	...	...	...	...	...	...
18270.0	111.95	171.20	283.15	141.575000	2	2011-03-18 12:41:00	2011-11-01 13:57:00	228	114.000000
18272.0	340.72	753.66	2710.70	542.140000	5	2011-04-07 09:35:00	2011-10-25 11:52:00	201	40.200000
18273.0	51.00	102.00	153.00	76.500000	2	2011-03-27 11:22:00	2011-09-05 11:27:00	162	81.000000
18283.0	1.95	313.65	1886.88	125.792000	15	2011-01-06 14:14:00	2011-11-30 12:59:00	327	21.800000
18287.0	70.68	1001.32	1837.28	612.426667	3	2011-05-22 10:39:00	2011-10-28 09:29:00	158	52.666667

Visualisation

1. 'Sale count' distribution
2. 'avg. number of days between purchases' distribution

Precdiction data set prepare

Aim: Predict M0: 1/1/2012 – 1/3/2012

1. M1: 30/09/2011 – 31/12/2011 (but the data set only end up 1/12, so two moth only)
2. M2: 30/06/2011 – 30/09/2011 (3 month)

3. M3: 31/3/2011 – 30/06/2011 (3 month)
4. M4: 31/12/2010 – 31/3/2011 (3 month)
5. M5: 1/12/2010 – 31/12/2010 (1 month)
6. ‘data_df’

	CustomerID	InvoiceDate	sales_sum	sales_avg	sales_count	M
0	12346.0	2011-03-31	77183.60	77183.600000	1	M_4
1	12347.0	2010-12-31	711.79	711.790000	1	M_5
2	12347.0	2011-03-31	475.39	475.390000	1	M_4
3	12347.0	2011-06-30	1018.77	509.385000	2	M_3
4	12347.0	2011-09-30	584.91	584.910000	1	M_2
...	...	...	...	...	...	...
9215	18283.0	2011-06-30	524.68	131.170000	4	M_3
9216	18283.0	2011-09-30	278.09	92.696667	3	M_2
9217	18283.0	2011-12-31	766.21	153.242000	5	M_1
9218	18287.0	2011-06-30	765.28	765.280000	1	M_3
9219	18287.0	2011-12-31	1072.00	536.000000	2	M_1

9220 rows × 6 columns

7. ‘Features_df’

CustomerID	sales_avg_M_2	sales_avg_M_3	sales_avg_M_4	sales_avg_M_5	sales_count_M_2	sales_count_M_3	sales_count_M_4	sales_count_M_5	sales_sum_M_2	sales_sum_M_3	sales_sum_M_4	sales_sum_M_5
12346.0	NaN	NaN	77183.600	NaN	NaN	NaN	1.0	NaN	NaN	NaN	77183.60	NaN
12347.0	584.91	509.385	475.390	711.79	1.0	2.0	1.0	1.0	584.91	1018.77	475.39	711.79
12348.0	310.00	367.000	227.440	892.80	1.0	1.0	1.0	1.0	310.00	367.00	227.44	892.80
12350.0	NaN	NaN	334.400	NaN	NaN	NaN	1.0	NaN	NaN	NaN	334.40	NaN
12352.0	316.25	NaN	312.362	NaN	2.0	NaN	5.0	NaN	632.50	NaN	1561.81	NaN

8. Response_df

	CustomerID	CLV_3M
5	12347.0	1294.32
10	12349.0	1757.55
14	12352.0	311.73
20	12356.0	58.35
21	12357.0	6207.67

9. **`Sample_set_df`**

	sales_avg_M_2	sales_avg_M_3	sales_avg_M_4	sales_avg_M_5	sales_count_M_2	sales_count_M_3	sales_count_M_4	sales_count_M_5	sales_sum_M_2	sales_sum_M_3	sales_sum_M_4	sales_sum_M_5	CustomerID	CLV_3M
NaN	0.00	0.000	77183.600	0.00	0.0	0.0	1.0	0.0	0.00	0.00	77183.60	0.00	12346.0	0.00
5.0	584.91	509.385	475.390	711.79	1.0	2.0	1.0	1.0	584.91	1018.77	475.39	711.79	12347.0	1294.32
NaN	310.00	367.000	227.440	892.80	1.0	1.0	1.0	1.0	310.00	367.00	227.44	892.80	12348.0	0.00
NaN	0.00	0.000	334.400	0.00	0.0	0.0	1.0	0.0	0.00	0.00	334.40	0.00	12350.0	0.00
14.0	316.25	0.000	312.362	0.00	2.0	0.0	5.0	0.0	632.50	0.00	1561.81	0.00	12352.0	311.73

Modeling Tranining

1. x_train : 0.7 * all features
2. x_test: 0.3 * all features
3. y_train: 0.7 * all target_var
4. y_test: 0.3 * all target_var
5. Result

	feature	coef
0	sales_avg_M_2	-0.319982
1	sales_avg_M_3	0.061492
2	sales_avg_M_4	0.244423
3	sales_avg_M_5	-0.460607
4	sales_count_M_2	46.839329
5	sales_count_M_3	-42.599988
6	sales_count_M_4	-27.110127
7	sales_count_M_5	54.269246
8	sales_sum_M_2	0.460445
9	sales_sum_M_3	0.368344
10	sales_sum_M_4	0.017492
11	sales_sum_M_5	0.481500

6. In-sample R squared: how well the model fits the training data
7. Out sample R squared: how well the model performs on testing data
8. In-sample MSE
9. Out sample MSE

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$