**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’

A screenshot of a data

AI-generated content may be incorrect.

1. ‘summary\_df’
2. A screenshot of a computer

   AI-generated content may be incorrect.

**Visualisation**

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

**Precdiction data set prepare**

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’

A screenshot of a data table

AI-generated content may be incorrect..

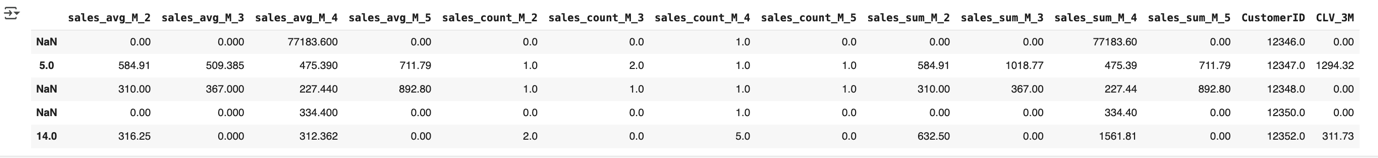
1. ‘Features\_df’A screenshot of a white sheet

   AI-generated content may be incorrect.
2. Response\_df

**A screenshot of a computer

AI-generated content may be incorrect.**

1. **Sample\_set\_df**



**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



1. In-sample R squared: how well the model fits the ​training data
2. Out sample R squared: how well the model performs on ​ testing data
3. In-sample MSE
4. Out sample MSE

A mathematical equation with numbers and symbols

AI-generated content may be incorrect.