**Data Preparation**

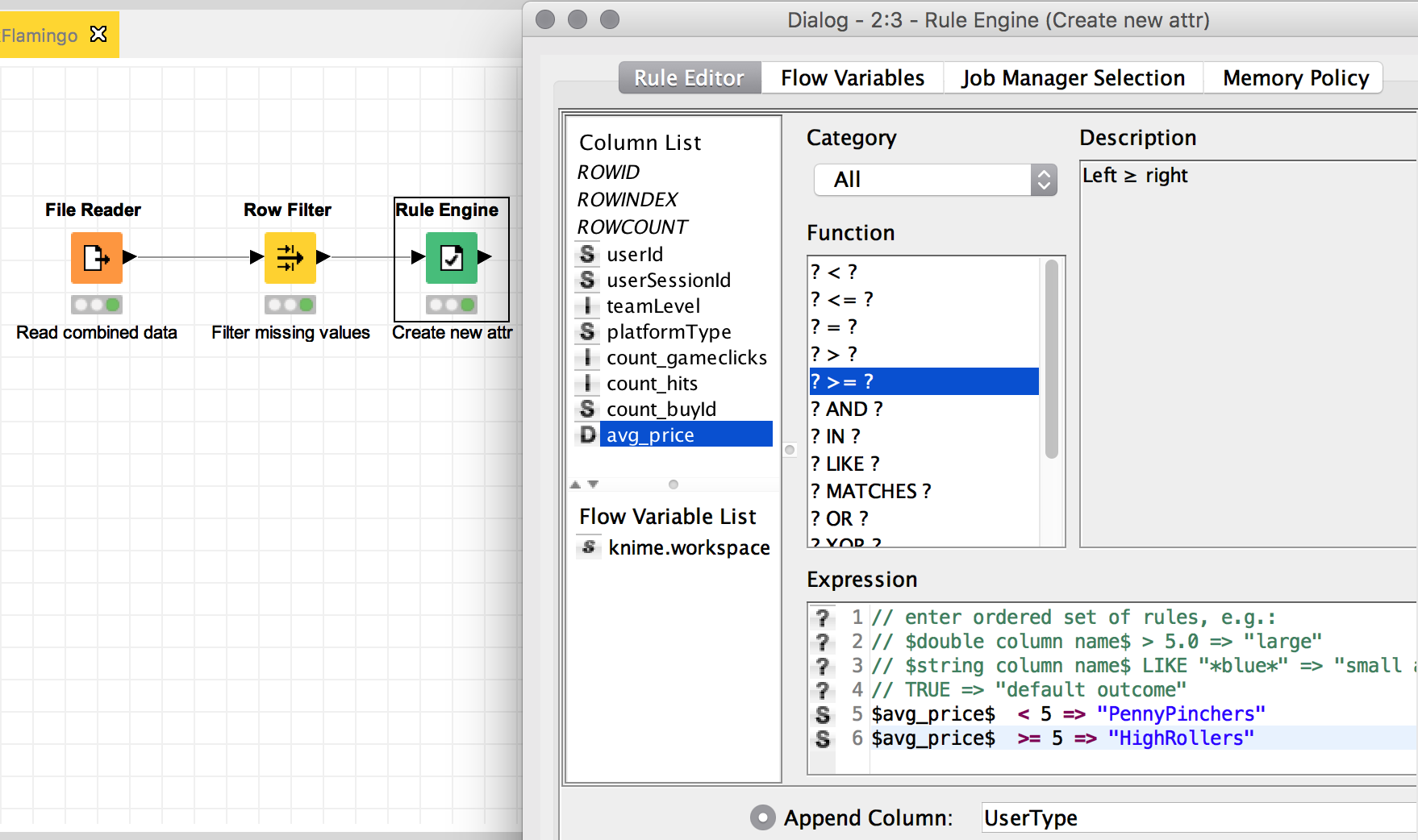
Analysis of combined\_data.csv

Sample Selection

|  |  |
| --- | --- |
| **Item** | **Amount** |
| # of Samples | 4,619 |
| # of Samples with Purchases | 1,411 |

Attribute Creation

A new categorical attribute was created to enable analysis of players as broken into 2 categories (HighRollers and PennyPinchers). A screenshot of the attribute follows:



A Rule Engine node was used for the creation of a new attribute call ‘User Type’, this node takes a rule and match them to each row in the table. If a rule matches, its outcome is added into the ‘UserType’ column. The rule used was:

**IF avg\_price < 5 THEN ‘PennyPinchers’**

**IF avg\_price >= 5 THEN ‘HighRollers’**

The creation of this new categorical attribute was necessary because we need to apply a classification algorithm, a decision tree works over a categorical variable, and we had to transform a quantitative variable into two categories.

Attribute Selection

The following attributes were filtered from the dataset for the following reasons:

|  |  |
| --- | --- |
| **Attribute** | **Rationale for Filtering** |
| userId | It’s an identifier and doesn’t give any information about the user and his likely to purchase big-tickets or not. |
| userSessionId | It’s an identifier and doesn’t give any information about the user and his likely to purchase big-tickets or not. |
| avg\_price | The information is now contained in the new attribute that we created in the step above. |