**Data Preparation**

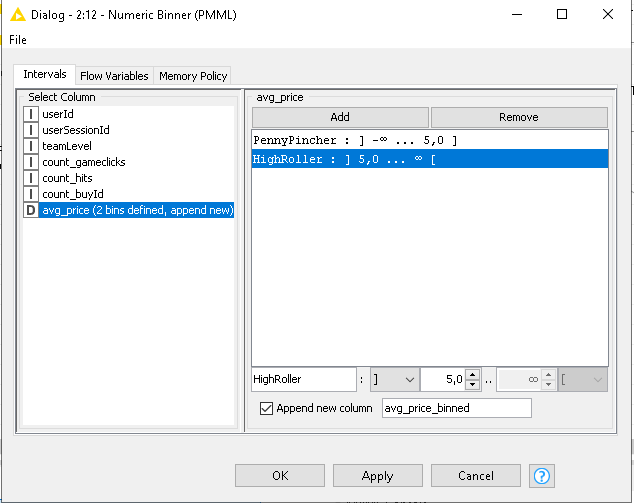
Analysis of combined\_data.csv

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:



If the avg\_price is lower than or equal to 5 dollar the new attribute is PennyPincher. If the avg\_price is higher than 5 dollar the new attribute is HighRoller.

The creation of this new categorical attribute was necessary because this compresses a numerical attribute with a wide range into a binary attribute. This reduction is beneficial when constructing the decision tree.

Attribute Selection

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

|  |  |
| --- | --- |
| **Attribute** | **Rationale for Filtering** |
| userID | This attribute is filtered since this is user specific. We wat to create a model applicable for all users, and thus also new users. |
| userSessionID | The userSessionID is a unique value for each session. This does not contribute in any way when making predictions. |
| avg\_price | This attribute is highly correlated to the binned version of this attribute. Using this when creating the model will result in a model with an extremely high accuracy based only on the avg\_price. |