**Data Preparation**

Analysis of combined\_data.csv

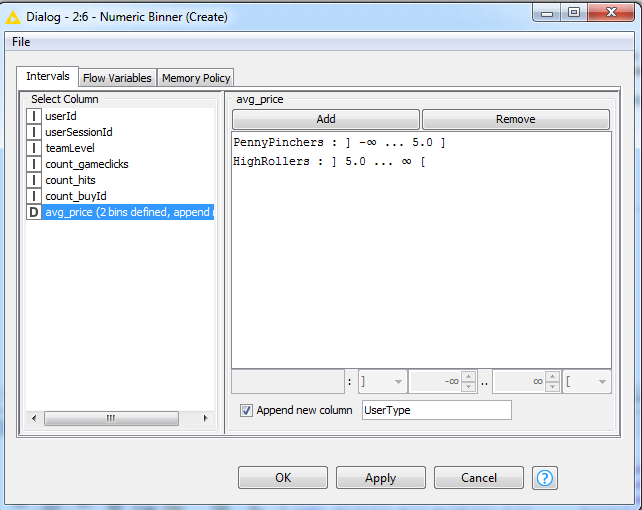
**Sample Selection**

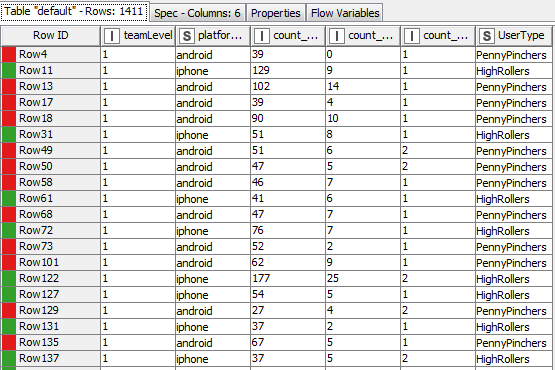
|  |  |
| --- | --- |
| **Item** | **Amount** |
| # of Samples | 4619 |
| # of Samples with Purchases | 1411 |

\*\*\*Number of samples without purchases = 3208

**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:





Describe the design of your attribute in 1-3 sentences.

* New column named **“UserType”** was added
* **PennyPinchers** have avg\_price <= 5.0$. Colored in Red (first bin)
* **HighRollers** have avg\_price > 5.0$. Colored in Green (second bin)

The creation of this new categorical attribute was necessary, because

* This new category is the **target variable** used for **data labelling of a classification task**. A classification task needs discrete categories
* For a **supervised learning classification task** such as our current task, labels are required during **model training**
* The **model score** is also derived from **comparing** **predicted labels & actual labels** of the **test set**

**Attribute Selection**

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

|  |  |
| --- | --- |
| **Attribute** | **Rationale for Filtering** |
| UserID | Assigned **Randomly** by system. Has no relationship to a user’s in-game behavior |
| SessionID | Assigned **Randomly** by system. Has no relationship to a user’s in-game behavior |
| Avg\_price | This is the column we **derive target variable from** 🡪 it has **100% correlation** to the target variable  We get rid of this feature because we already have the **UserType column** which acts as **labels for the classification task** |

\*Remaining features = Team\_level, platformType, count\_clicks, count\_ishits, count\_buyid