## Data Preparation

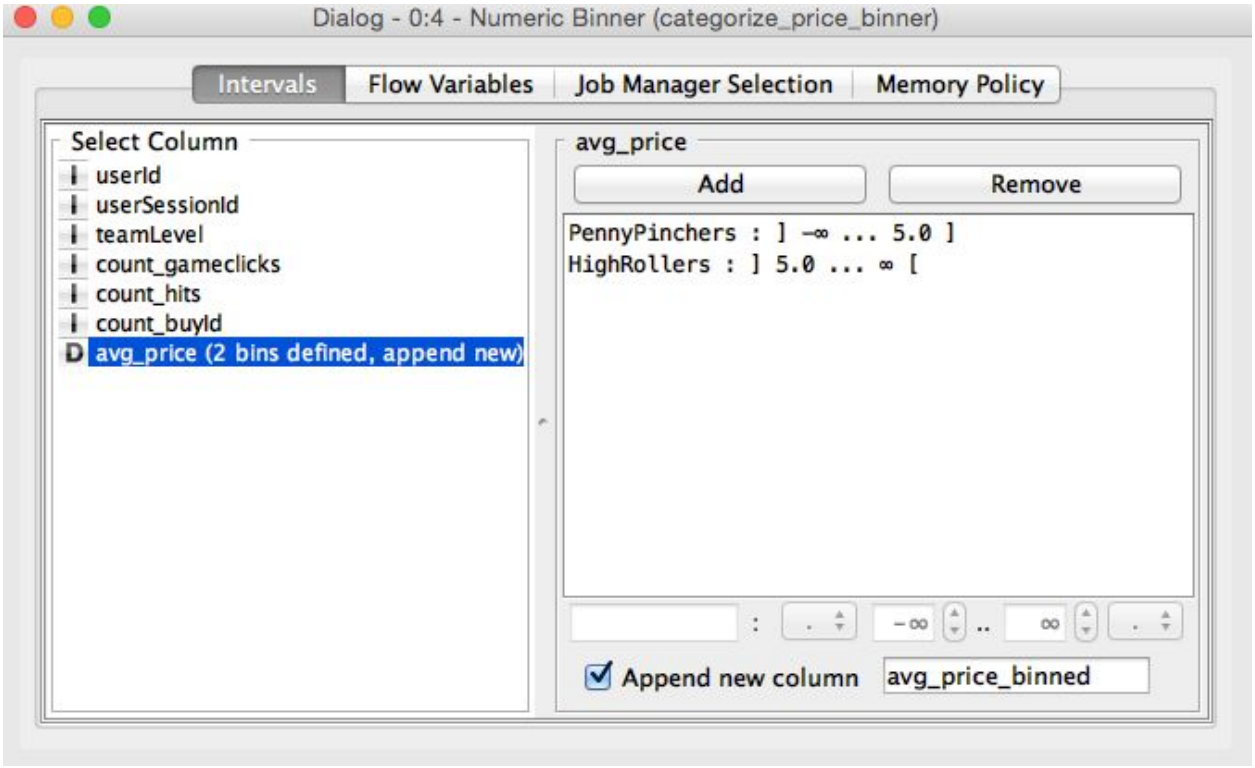
Analysis of combined_data.csv

Sample Selection

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

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:




The categorical attribute which I have named *avg_price_binned* will be used as the target during the classification process to categorize the users under two classes: buyers of items that cost more than $5.00 - shown above as HighRollers - and buyers of items that cost $5.00 or less - shown above as PennyPinchers. The table above maps any average price from negative infinity (although a negative price does not make sense so we can assume an item average price start

at $0) to an average price of $5.0 including $5.0 as PennyPinchers. The “]” bracket after 5.0 means that the number 5 is included in the PennyPinchers category. Whereas, HighRollers include any average price that is bigger than 5.0. The “]” bracket in this case implies that the value 5.0 is excluded from the HighRollers class. The *avg_price_binned* is a binary categorical variable of either 0/1 or True/False.

The creation of this new categorical attribute was necessary because it will group user behavior under two categories and will allow us to predict with some level of uncertainty whether new users would fall under PennyPinchers or HighRollers. This information is also useful for targeting users who are HighRollers with ads of items that are more expensive which could increase revenues from items.

Attribute Selection

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

| **Attribute** | **Rationale for Filtering** |
|---|---|
| avg_price | Since this is the information we are trying to predict as a categorical variable, it should not be included in the classification |
| userId | Since we are trying to predict the users that are more likely to buy items that cost more than $5.0, userId does not affect this |
| userSessionId | userSessionId does not add any knowledge or value to finding which users are more likely to buy items that cost more than $5.0 |
| | |