2.1 Data Preparation

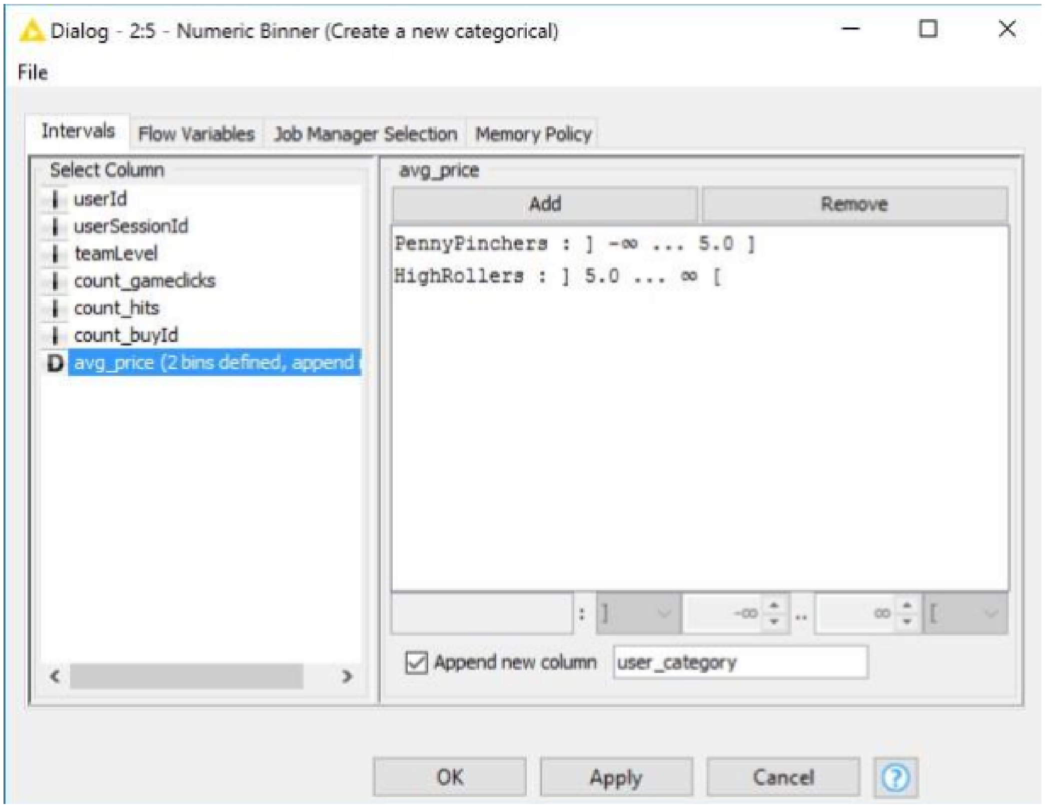
Analysis of combined_data.csv to investigate the classification of users.

2.1.1 Sample Selection

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

2.1.2 Attribute Creation

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



Dialog - 2:5 - Numeric Binner (Create a new categorical)

Select Column: userId, userSessionId, teamLevel, count_gameclicks, count_hits, count_buyId, avg_price (2 bins defined, append

avg_price

```
PennyPinchers : ] -∞ ... 5.0 ]
HighRollers : ] 5.0 ... ∞ [
```

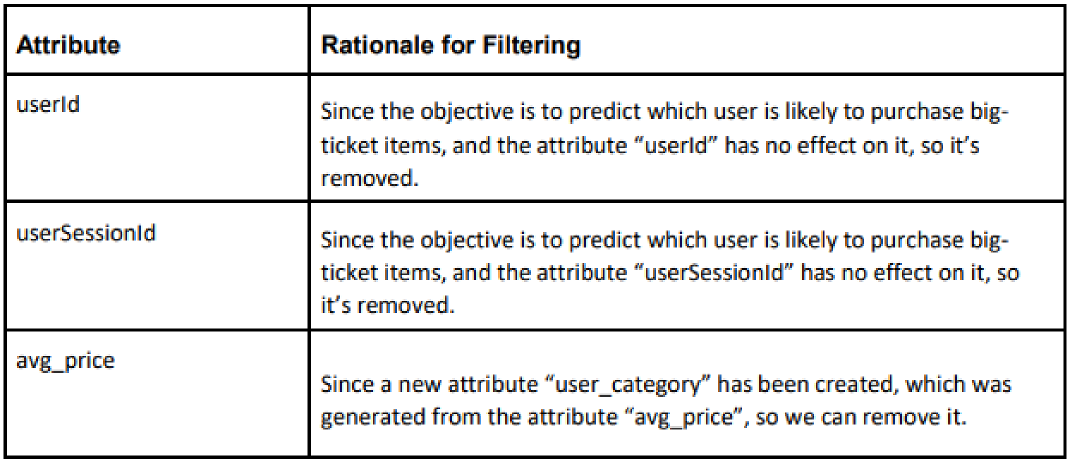
☑ Append new column: user_category

A new categorical attribute, named “user_category”, is created by the Numeric Binner node. As presented in the instruction, we need to define two categories for price which we will use to distinguish between HighRollers(buyers of items that cost more than $5.00) and PennyPinchers (buyers of items that cost $5.00 or less), so as we see in the screenshot above, the user who costs $5.00 or less is defined as “PennyPinchers”, the user who costs more than $5.00 is defined as “HighRollers”.

The creation of this new categorical attribute was necessary because it can facilitate the classification of users and contribute to the following steps.

2.1.3 Attribute Selection

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



Attribute	**Rationale for Filtering**
userId	Since the objective is to predict which user is likely to purchase big-ticket items, and the attribute “userId” has no effect on it, so it’s removed.
userSessionId	Since the objective is to predict which user is likely to purchase big-ticket items, and the attribute “userSessionId” has no effect on it, so it’s removed.
avg_price	Since a new attribute “user_category” has been created, which was generated from the attribute “avg_price”, so we can remove it.