**Data Preparation**

Analysis of combined\_data.csv

Nyan Lynn Htet | 8 Sept 2024

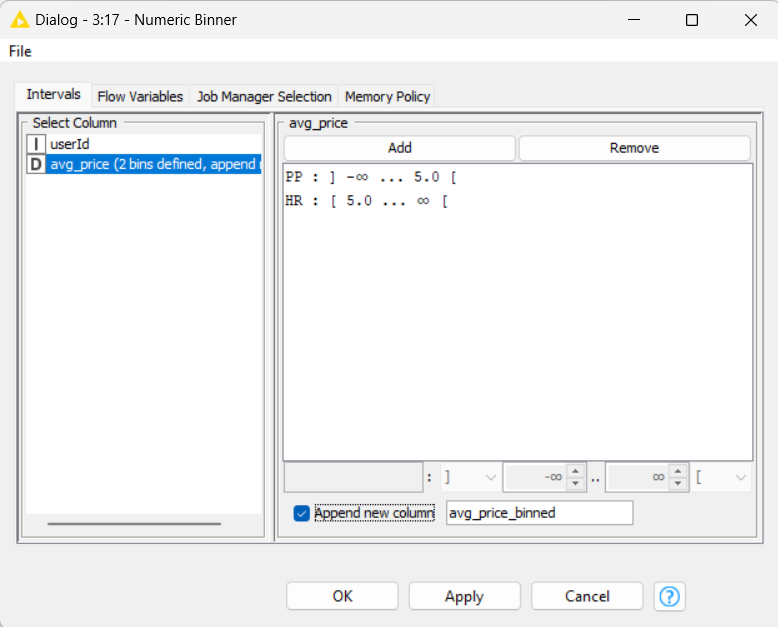
**Week2 Assignment: Classifying in KNIME (Catch the Pink Flamingo)**

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:



*Note\* HR = HighRollers, PP = PennyPinchers*

**Description**: HighRollers were defined those who spent $5.0 (on average price) and over and PennyPinchers are those who spent less than $5.0 (on average price).

The creation of this new categorical attribute was necessary because:

We aim to identify the group of people who spend generous amount and those who spend less. Then make a marketing strategy for the game according to our findings so that we can raise the sales. Also help in the decision making process.

Attribute Selection

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

|  |  |
| --- | --- |
| **Attribute** | **Rationale for Filtering** |
| count\_buyId | To filter out NULL since their purchase data is not available |
| userSessionId | Not relevant for the model this time |
| avg\_price | To sort this so that we identify the HighRoller and PennyPinchers |