**Kedir Nasir Omer Date: Nov. 28/2022**

**UC SanDiego- Big Data Specialization**

**2.1. Data Preparation**

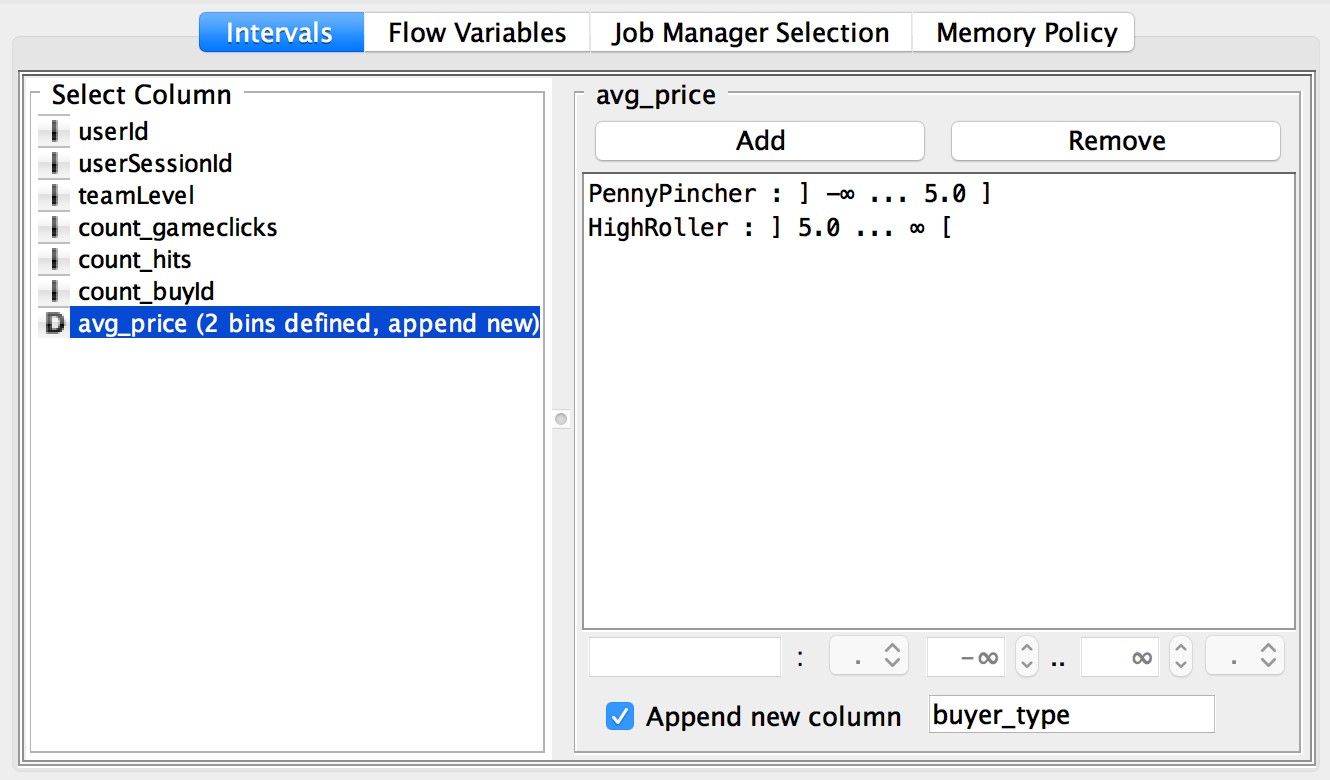
* **Classifying in KNIME to identify big spenders in Catch the Pink Flamingo**
* Analyzing the combined\_data.csv

# Sample Selection

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

# Attribute Creation

For the purpose of categorizing players into two groups for analysis, a new categorical attribute was developed (**HighRollers** and **PennyPinchers**).



* **Describing the design attribute**

High rollers are those who make purchases that cost more than $5. We can categorize users in the appropriate way by creating a new column based on the avg price.

Because our objective is to comprehend the characteristics of those who make significant purchases, the development of this new categorical attribute was necessary. We will build our decision tree around this categorical variable.

# Attribute Selection

For the reasons listed below, the following attributes were excluded from the dataset

|  |  |
| --- | --- |
| **Attribute** | **Rationale for Filtering** |
| userId | Not relevant for the model. |
| userSessionId | Not relevant for model. |
| avg\_price | This feature was used to create the categorical feature “buyer\_type”, the variable we are trying to predict based on other elements. We do not want to include this in our model. |