# Data Preparation

*Coursera Big Data Specialization Capstone Project, Week 2*

**Peer Graded Assignment: Classifying in KNIME to identify big spenders in Catch the Pink Flamingo**

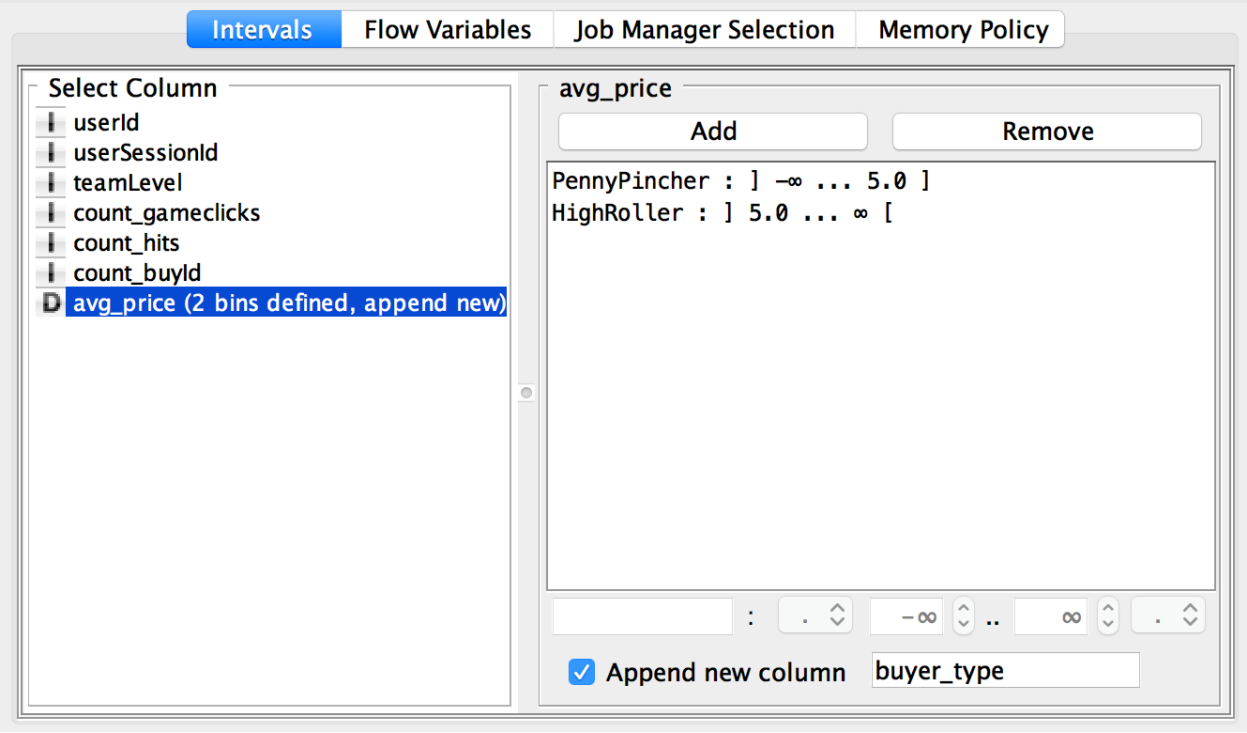
Analysis of combined_data.csv

## 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:



**Describe the design of your attribute in 1-3 sentences:**
High rollers are defined as those who purchase items over $5.00. Defining a new column based on the avg_price allows us to classify users accordingly.

The creation of this new categorical attribute was necessary because **our goal is to understand the attributes of who makes large purchases. This categorical variable is what we are going to base our decision tree upon.**

### Attribute Selection

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

| **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. |