Data Preparation

Analysis of combined_data.csv

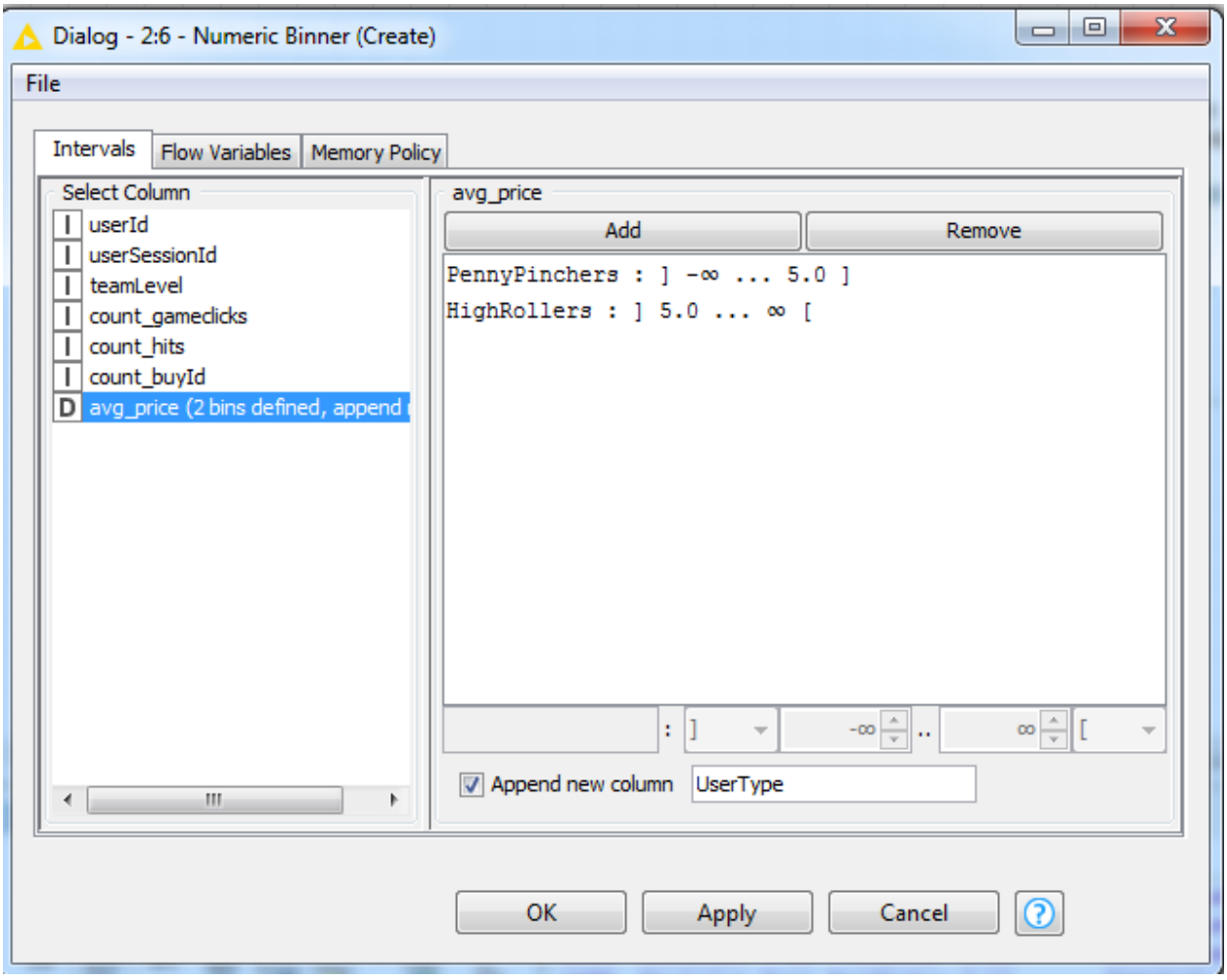
Sample Selection

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

***Number of samples without purchases = 3208

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:



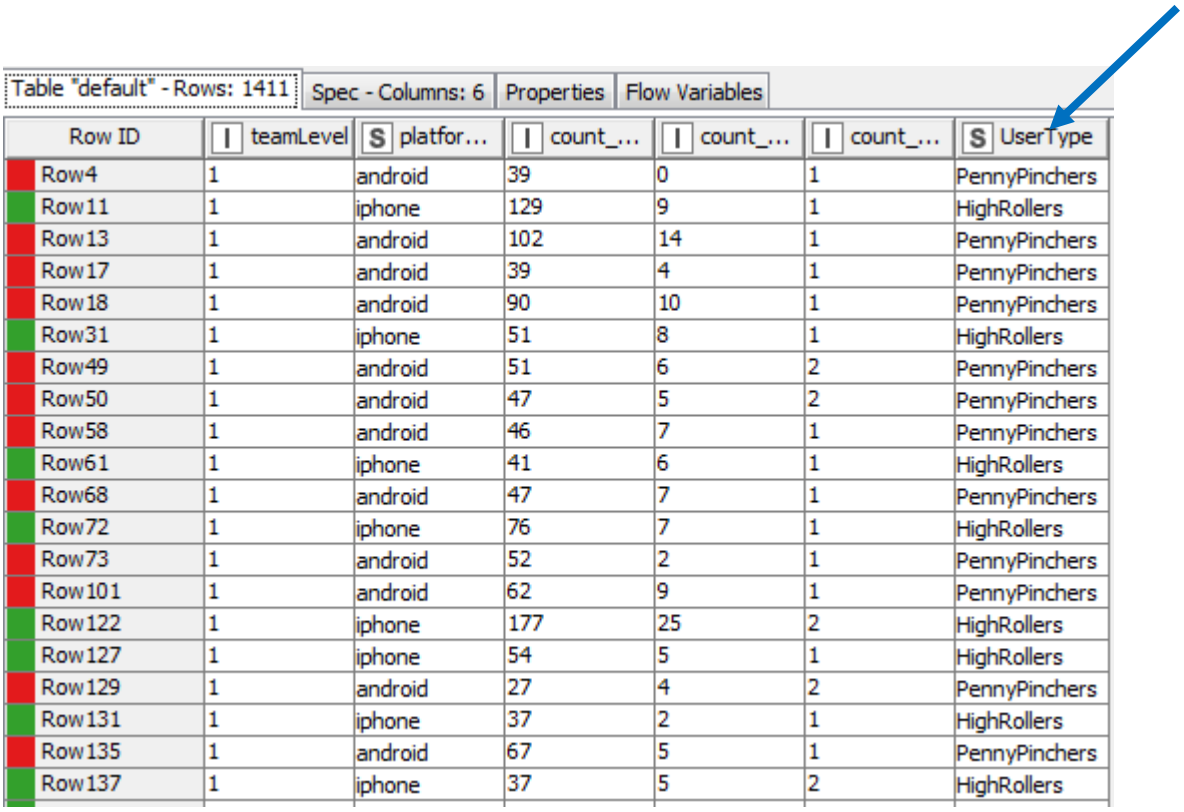


Table "default" - Rows: 1411 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	teamLevel	platfor...	count_...	count_...	count_...	UserType
Row4	1	android	39	0	1	PennyPinchers
Row11	1	iphone	129	9	1	HighRollers
Row13	1	android	102	14	1	PennyPinchers
Row17	1	android	39	4	1	PennyPinchers
Row18	1	android	90	10	1	PennyPinchers
Row31	1	iphone	51	8	1	HighRollers
Row49	1	android	51	6	2	PennyPinchers
Row50	1	android	47	5	2	PennyPinchers
Row58	1	android	46	7	1	PennyPinchers
Row61	1	iphone	41	6	1	HighRollers
Row68	1	android	47	7	1	PennyPinchers
Row72	1	iphone	76	7	1	HighRollers
Row73	1	android	52	2	1	PennyPinchers
Row101	1	android	62	9	1	PennyPinchers
Row122	1	iphone	177	25	2	HighRollers
Row127	1	iphone	54	5	1	HighRollers
Row129	1	android	27	4	2	PennyPinchers
Row131	1	iphone	37	2	1	HighRollers
Row135	1	android	67	5	1	PennyPinchers
Row137	1	iphone	37	5	2	HighRollers

Describe the design of your attribute in 1-3 sentences.

- New column named **“UserType”** was added
- **PennyPinchers** have avg_price <= 5.0$. Colored in Red (first bin)
- **HighRollers** have avg_price > 5.0$. Colored in Green (second bin)

The creation of this new categorical attribute was necessary, because

- This new category is the **target variable** used for **data labelling of a classification task**. A classification task needs discrete categories
- For a **supervised learning classification task** such as our current task, labels are required during **model training**
- The **model score** is also derived from **comparing predicted labels & actual labels** of the **test set**

Attribute Selection

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

Attribute	**Rationale for Filtering**
UserID	Assigned **Randomly** by system. Has no relationship to a user's in-game behavior
SessionID	Assigned **Randomly** by system. Has no relationship to a user's in-game behavior
Avg_price	This is the column we **derive target variable from** → it has **100% correlation** to the target variable We get rid of this feature because we already have the **UserType column** which acts as **labels for the classification task**

*Remaining features = Team_level, platformType, count_clicks, count_ishits, count_buyid