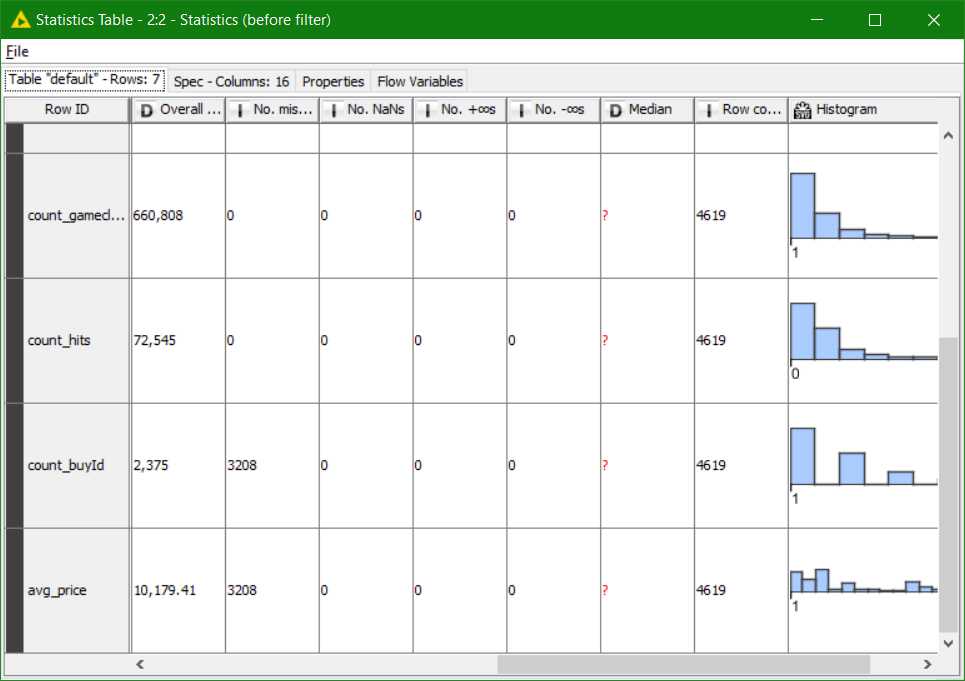
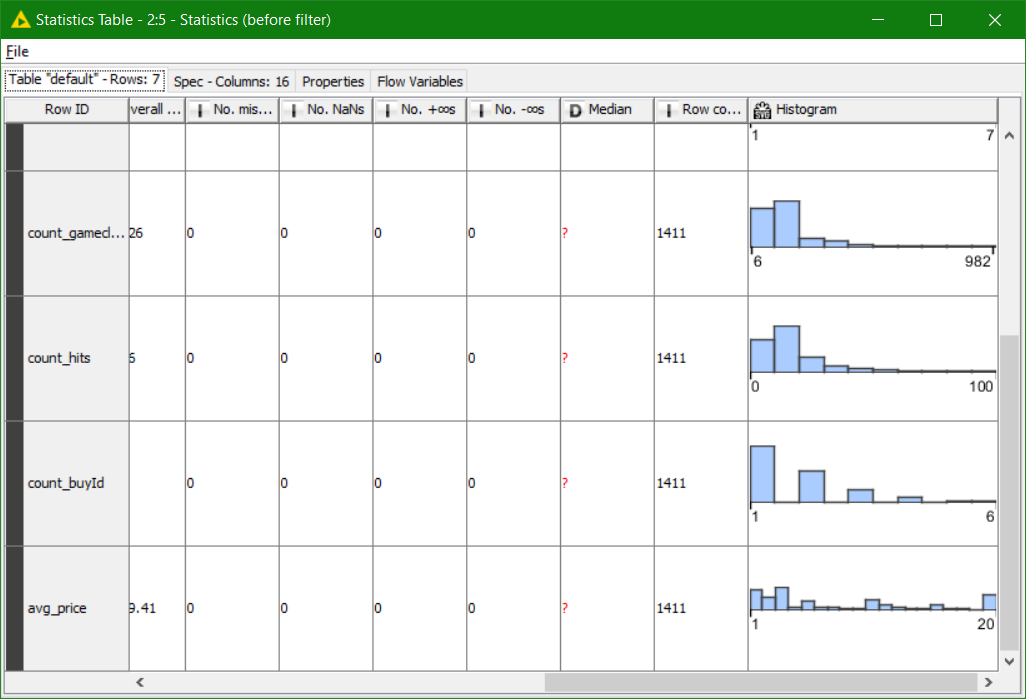
**Data Preparation**

Analysis of combined\_data.csv

Sample Selection

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

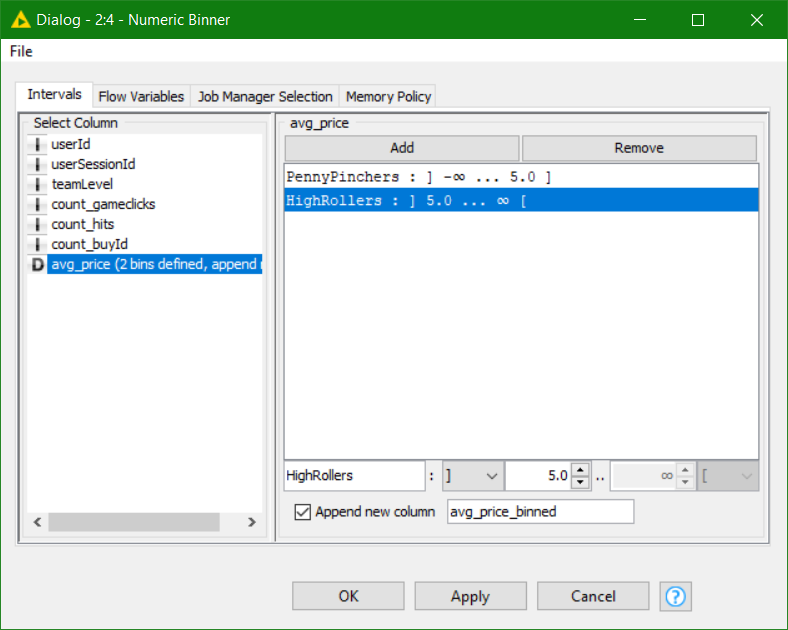


v

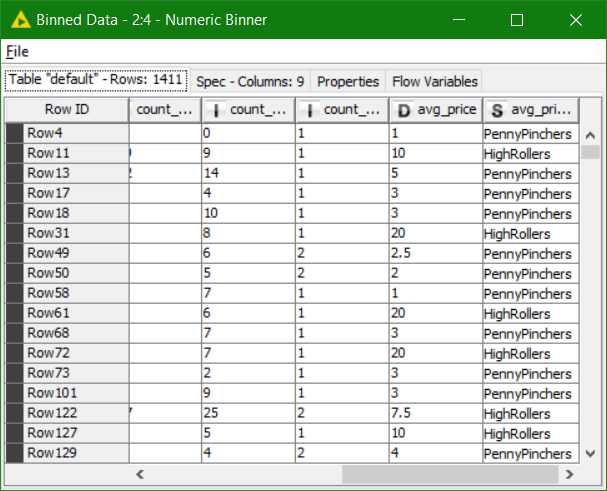
Here I have taken two snaps. First one is before applying filters (**4619 rows**) and the second one is after applying filters. Those rows which have NULL as a value are removed from the combined\_data.csv file and then the statistics are observed similar to the first. After filtering we get **1411** rows and also the graph is quite similar to the previous one.

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:



The numeric avg\_price variable was redefined as a category variable with 2 values: PennyPinchers and HighRollers. Penny Pinchers were those who bought items costing $5.00 or less, and HighRollers are those users who bought items costing above $5.00. The design is shown above, **where “]” is inclusive**, and **“[“ is exclusive**. The new category variable is named **“avg\_price\_binned.”**



The creation of this new categorical attribute was necessary because we will be using a decision tree algo. to determine the attributes responsible for **highroller** and **a pennyPincher**.

It will also serve as the reference for training and subsequently scoring the Decision Tree model.

Avg-price only can not be used for classification task with a continuous-value.

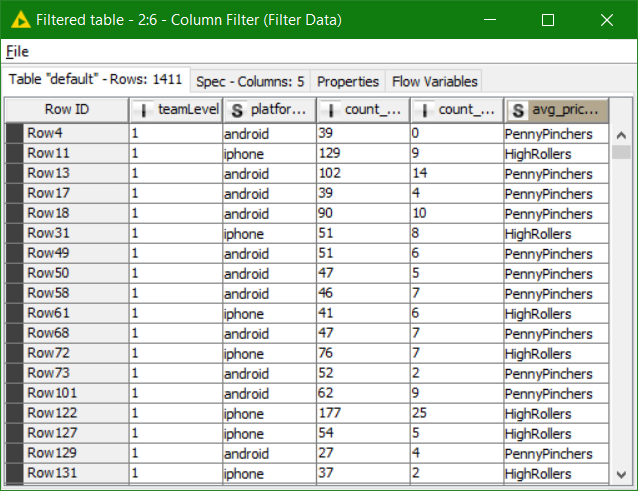
**Attribute Selection**

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

|  |  |
| --- | --- |
| **Attribute** | **Rationale for Filtering** |
| userId | We are not interested in finding who exactly is a highRoller plus It doesn’t make any sense on deciding whether a player is highRoller or a PennyPencher. |
| userSessionId | This attribute is used to identify the session and the session does not contribute to a highRoller or a PennyPincher |
| count\_buyId | The number of items purchased doesn’t define a highRoller |
| avg\_price | The numeric variable was replaced by the category variable(Binned). So it cannot be included within the training and testing dataset as the decision tree will then be giving us 100% Accuracy which defeat the original objective of analysing the behaviour and predicting the results. |

After applying column filter.

Now if you observe the last column, their you can find that it has specified the PennyPinchers and the highRollers



The resulting table will be passed to the Color Manager node, where **High Rollers will be assigned a Red color** and **PennyPinchers a Blue.**

