**Preliminary feature engineering and preliminary feature selection. Splitting data into training and testing sets, decision making and description of how model was trained.**

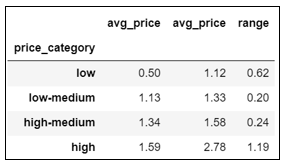
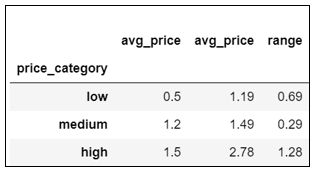
Our y-variable, that is avocado average price is ***continuous variable.*** We decided to use both approaches Regression and Classification. In order to use classification model, y-variable has to be ***categorical value***. We used **qcut function** to evenly distribute values into 3 and 4 categories. One advantage of qcut function is that values are evenly distributed; yet on the other hand, the range between categories might not be even. We kept this in mind when analyzing final results in confusion matrix.

We split data into training and testing at 75% and 25% respectively. Next, we run the models on various combinations of the X variables. For example:

* Using only single dataset “prices” (without production dataset) in order to see how the models perform only with prices dataset.
* Using ratio between total volume sold and total volume produced and without.
* Using only one type of the avocado at the time (conventional or organic) and combined.
* Removing noise (larger regions vs only cities and vice versa).
* Using only weeks, months and years individually and all at once.

When deciding about our final model for this dataset we looked at model performance scores, feature importances, and our final decision what features are important for our project (for examples cities vs regions). Our decision making includes:

* Model slightly improved when using combined dataset (prices and production); therefore, we decided to train the model using both datasets.
* Model performed better when using only regions and no cities; however, we decided to use the model only with the cities, since predicting prices in selected city is our focus.
* Ratios (production vs. volume sold) contributed to improved ML model, so we used those ratios as one of our X-variable.
* y-variable was split into 3 and 4 categories. Model performed better when split into 3 categories vs 4 categories. At first, we thought it might be better if we split y-variables into 4 categories in order to have more precise predictions. However, when checking bins’ ranges the “low-medium” and “high-medium” had the lowest and uneven range of prices. Based on that we decided to use 3 categories instead of 4.



*Figure 1: Bins ranges with qcut for 3 and 4 categories.*

* Using only weeks, months and years all at once showed the best results, so we decided to keep all three as our features.