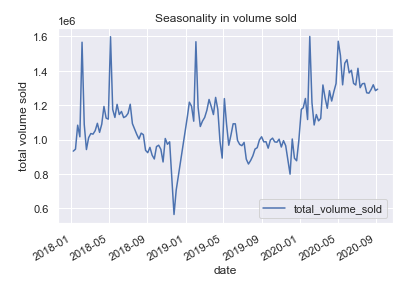
**Random Forest Ensemble (Regressor and Classifier) Preprocessing Steps**

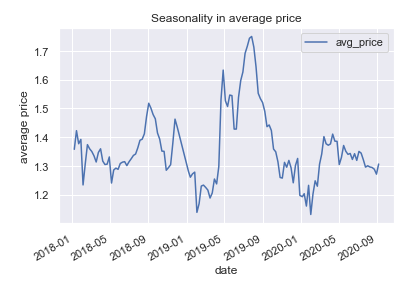
**Dataset**

We trained Random Forest Ensemble model on prices and production dataset.

* **y-variable** (target value) is the average price of the avocados in the US cities.
* **X-variables** are week, month, year, units sold (3 different types), bags sold (3 different types), production data for avocados in California, Chile, Peru, Mexico, Columbia and the ratio between total volume sold and total volume produced. This step was taken to reflect the ratio per each city, since the production data was repetitive per city. Using ratio in ML model showed improvement on model performance.

**Preliminary data preprocessing**

From our EDA we recognized seasonality in dataset, where we saw patterns in weeks and months for each year. This observation played big role in our decision making when building ML models and dealing with the time dimension. We approached to this problem, by splitting the date into weeks, months and years and use those features in ML model as separate X-variables.



*Figure 1 & 2: Seasonality in average price and volume sold.*

Additionally, our dataset had some outliers. These datapoints were combined regions that contained sums of selected cities. We recognized those as a noise in the dataset and decided to remove them from the dataset, since our focus in the project is selected cities in the US. Our data preprocessing includes:

* Converting date datatype from object to datetime and extracting week, month and year with Pandas datetime function.
* Making necessary calculations (ratios).
* Drop NaN values.
* Renaming confusing columns.
* Handling outliers and noise in the dataset (excluding regions and keeping only cities).
* Cleaning column “type” (some datapoints “conventional” had space at the end, so the value appeared twice “conventional” and “conventional ”).
* Encoding strings and dropping non-beneficial columns.
* Normalizing data set, using Standard Scaler.