**AVOCADO DATASET**

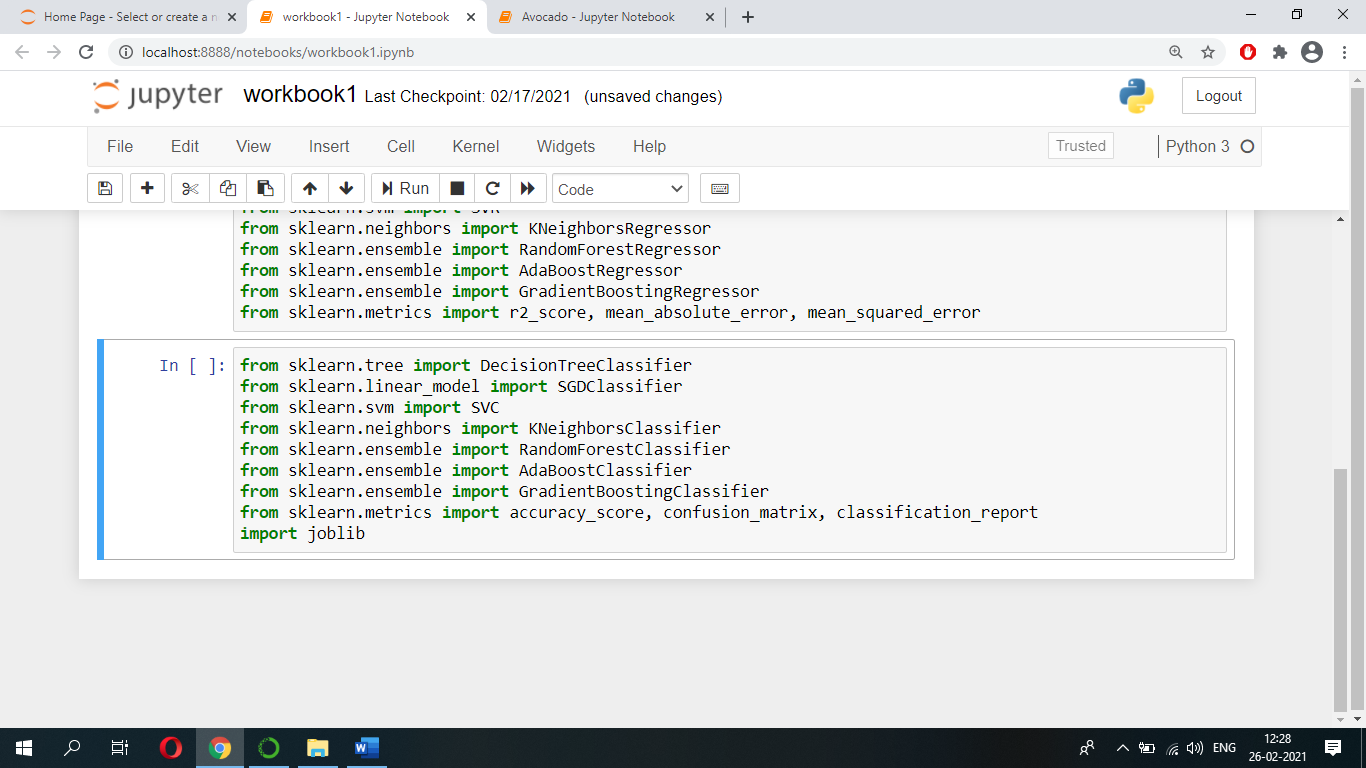
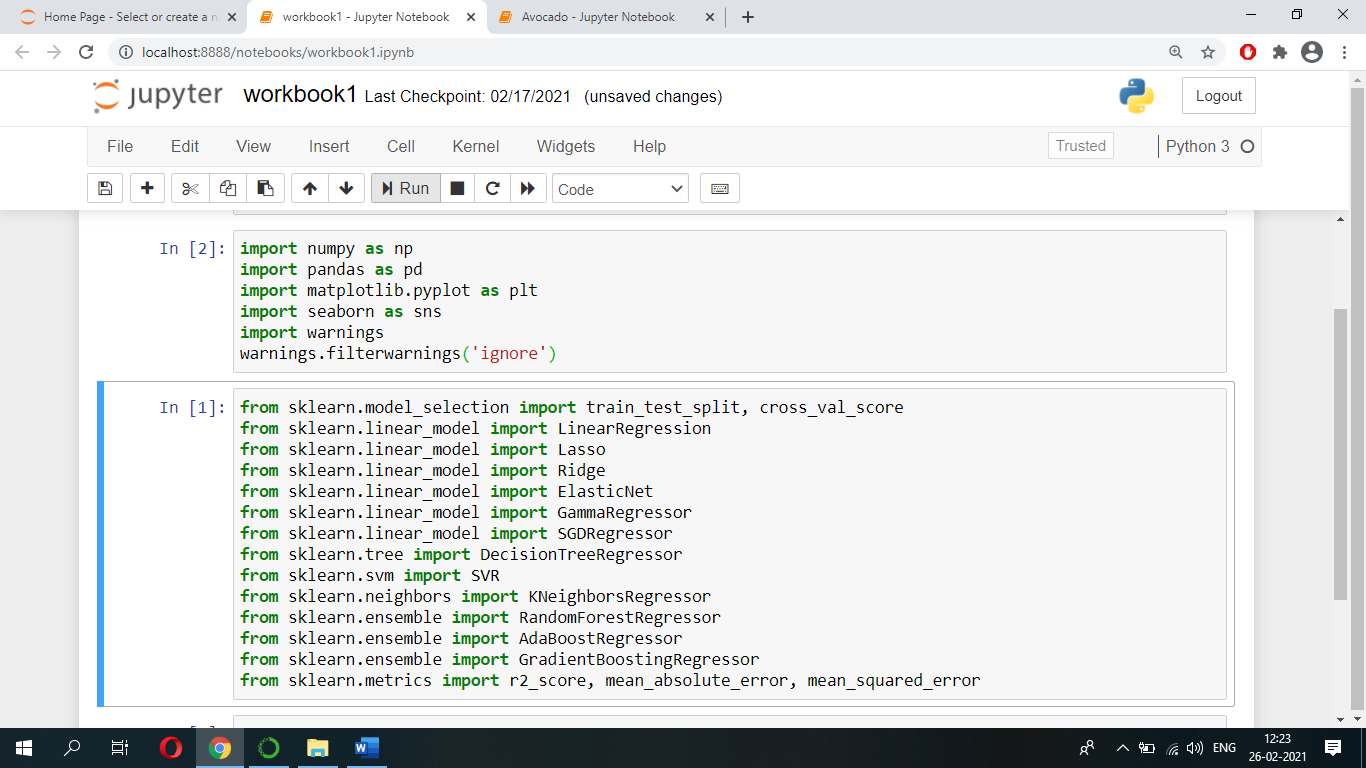
* **Introduction**

Avocado is a fruit consumed by people heavily in the United States.

This data was downloaded from the Hass Avocado Board website in May of 2018 & compiled into a single CSV. The table below represents weekly 2018 retail scan data for National retail volume (units) and price. Retail scan data comes directly from retailers’ cash registers based on actual retail sales of Hass avocados. The Average Price (of avocados) in the table reflects a per unit (per avocado) cost, even when multiple units (avocados) are sold in bags. The Product Lookup codes (PLU’s) in the table are only for Hass avocados. Other varieties of avocados (E.g., Green skins) are not included in this table.

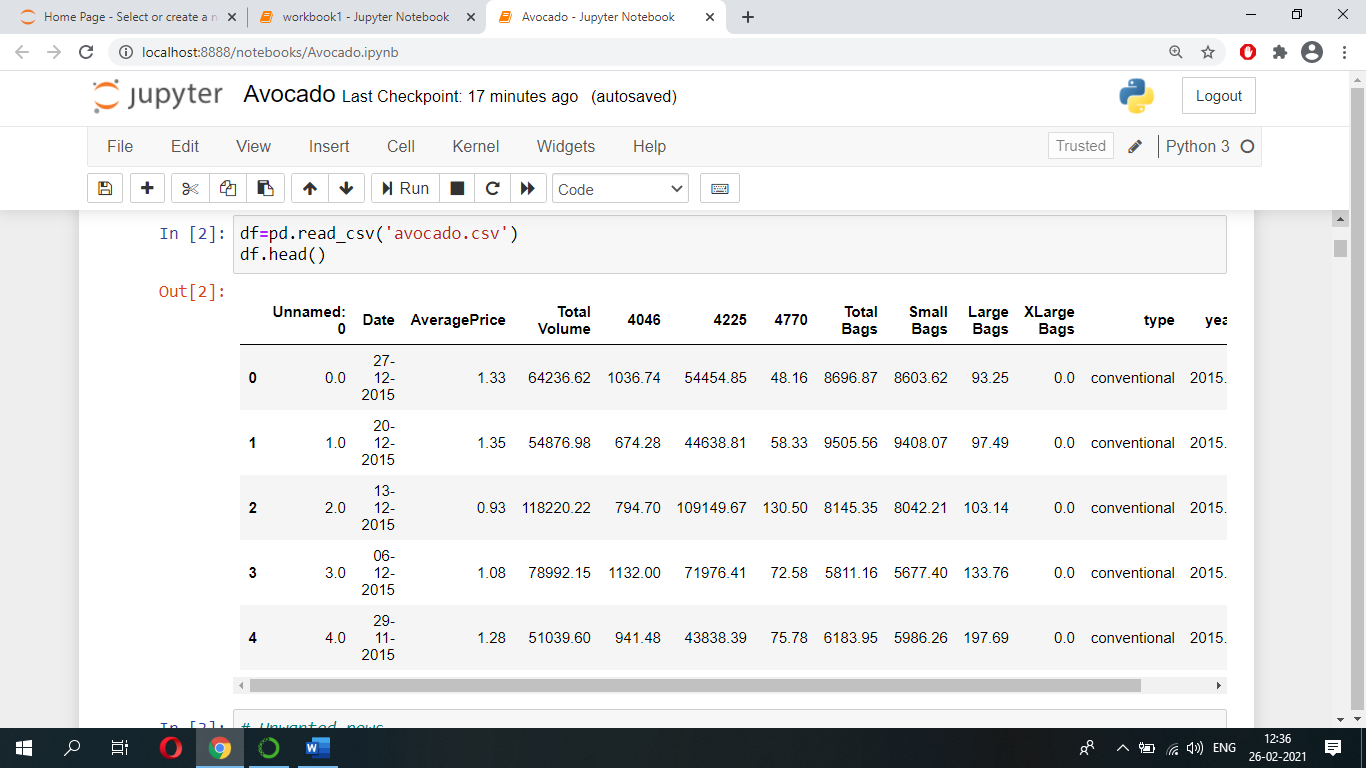
* **Problem Statement**
* To train a model/algorithm which can predict the average price of an avocado (Regression problem).
* To train a model/algorithm which can predict city / region (Classification problem).
* **Importing Libraries**

First of all, we will import ‘pandas’ to read our file from a csv file and manipulate it for further use. We will use ‘numpy’ to convert our data into a format suitable to feed our models. We will use ‘matplotlib’ and ‘seaborn’ for visualizations. We will then import models and its necessary metrics from ‘sklearn’. Then, we will import ‘joblib’ available in ‘sklearn’ to save our model for future use.

****

* **Data Description**

We have our data saved in a csv file named ‘avocado.csv’. We first read our dataset into a pandas dataframe called ‘df’, and then use the head() function to show the first five records of our dataset.



The following are the features/attributes present in our dataset:

1. Date - The date of observation
2. AveragePrice - the average price of a single avocado
3. Total Volume - Total number of avocados sold
4. 4046 - Total number of avocados with PLU 4046 sold (Product Lookup codes (PLU’s)
5. 4225 - Total number of avocados with PLU 4225 sold (Product Lookup codes (PLU’s)
6. 4770 - Total number of avocados with PLU 4770 sold (Product Lookup codes (PLU’s)

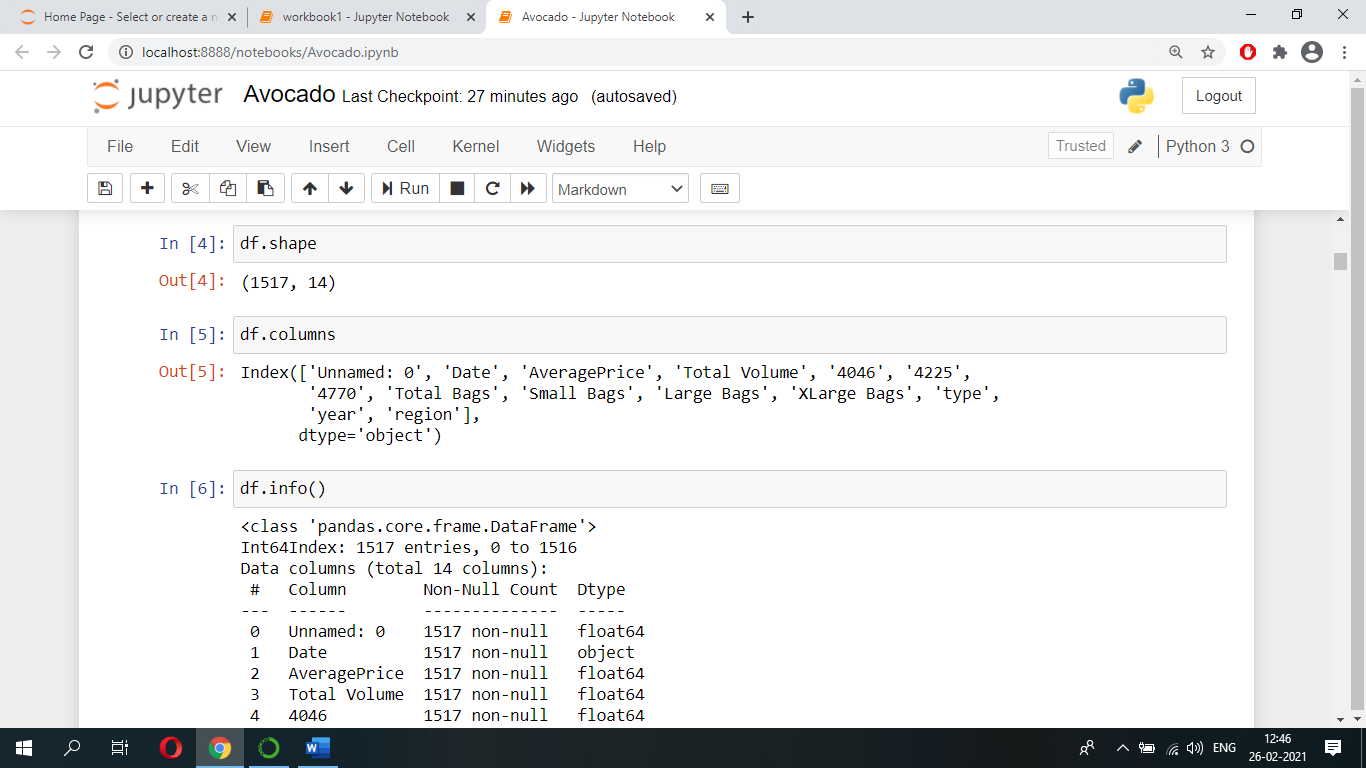
Note: The PLU code identifies produce items based upon the commodity, variety and size group.

1. Total Bags – How many Total bags of avocados are sold across US
2. Small Bags - How many Small bags of avocados are sold across US
3. Large Bags - How many Large bags of avocados are sold across US
4. XLarge Bags - How many XLarge bags of avocados are sold across US
5. Type - conventional / organic
6. Year - the year of observation
7. Region - the city or region of observation

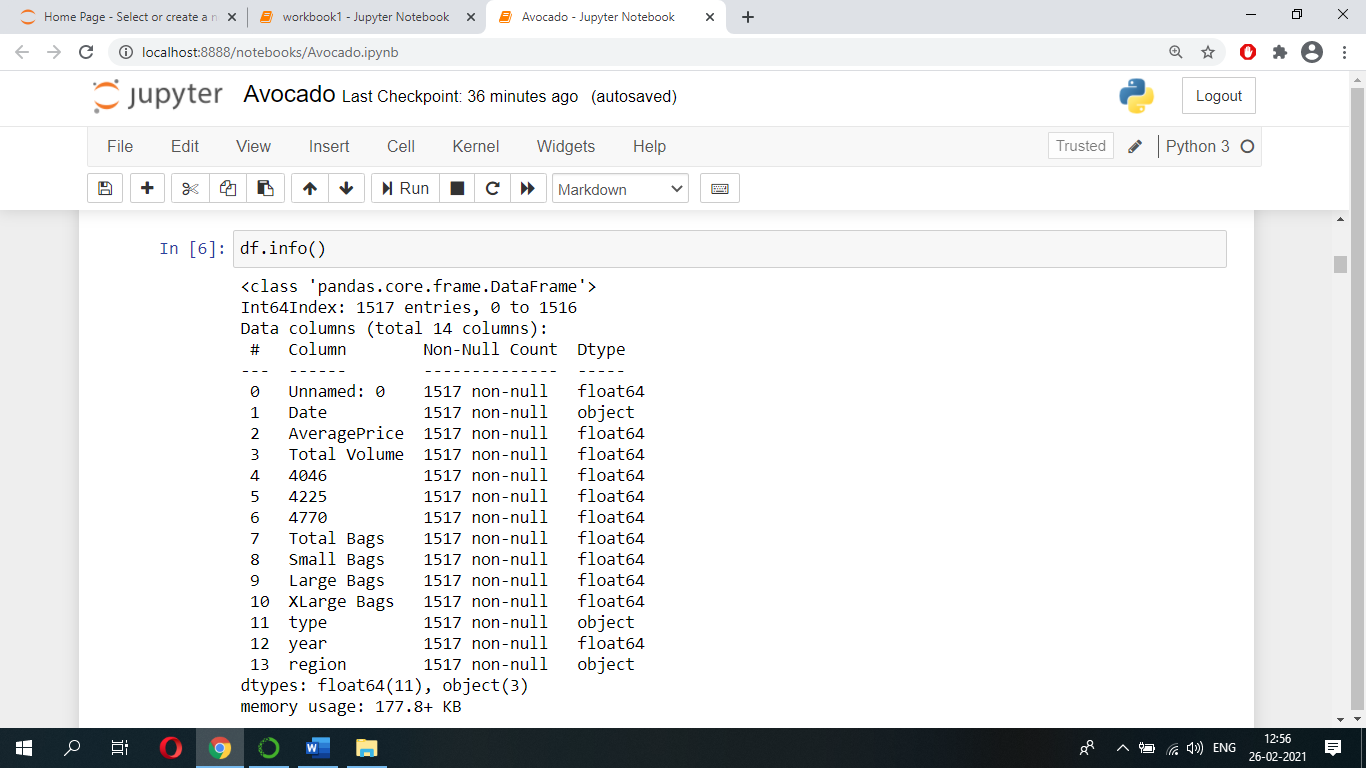
* **Exploratory Data Analysis (EDA)**

Let us now explore our dataset to get a feel of what it looks like and get some insights about it.

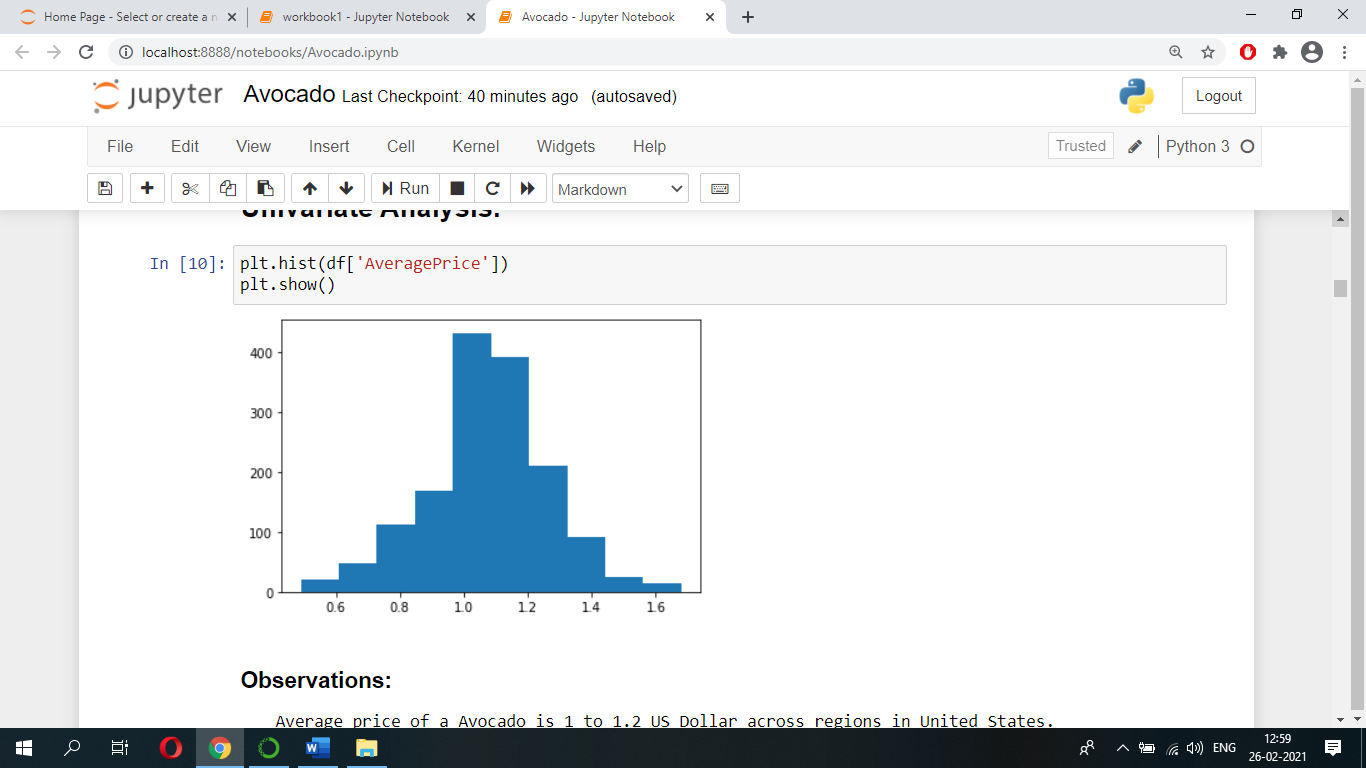
Let’s start by finding how many instances/rows we got in our dataset and which are the attributes/columns present in our dataset using shape and columns function.

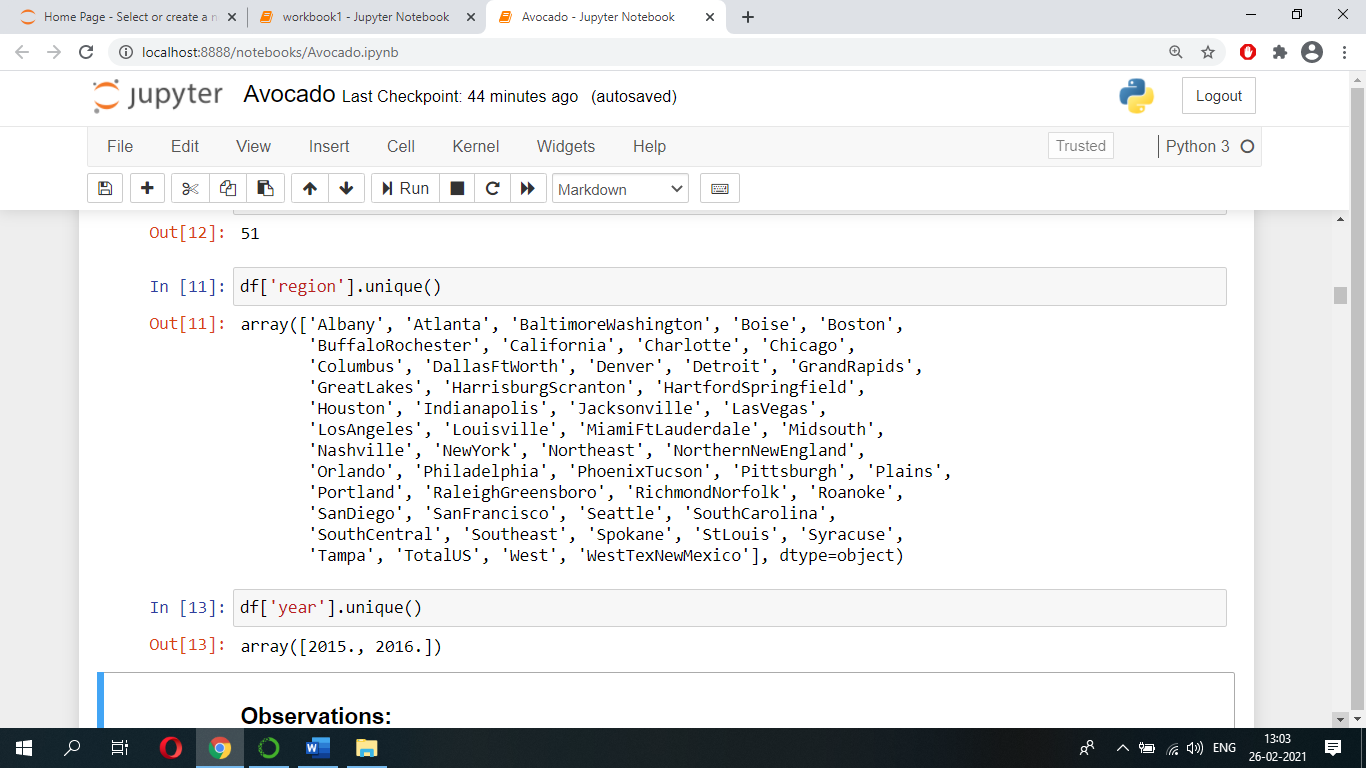


From above we can see that we got 1517 instances/rows and 14 attributes/columns. And we can see which are the attributes present in our dataset.

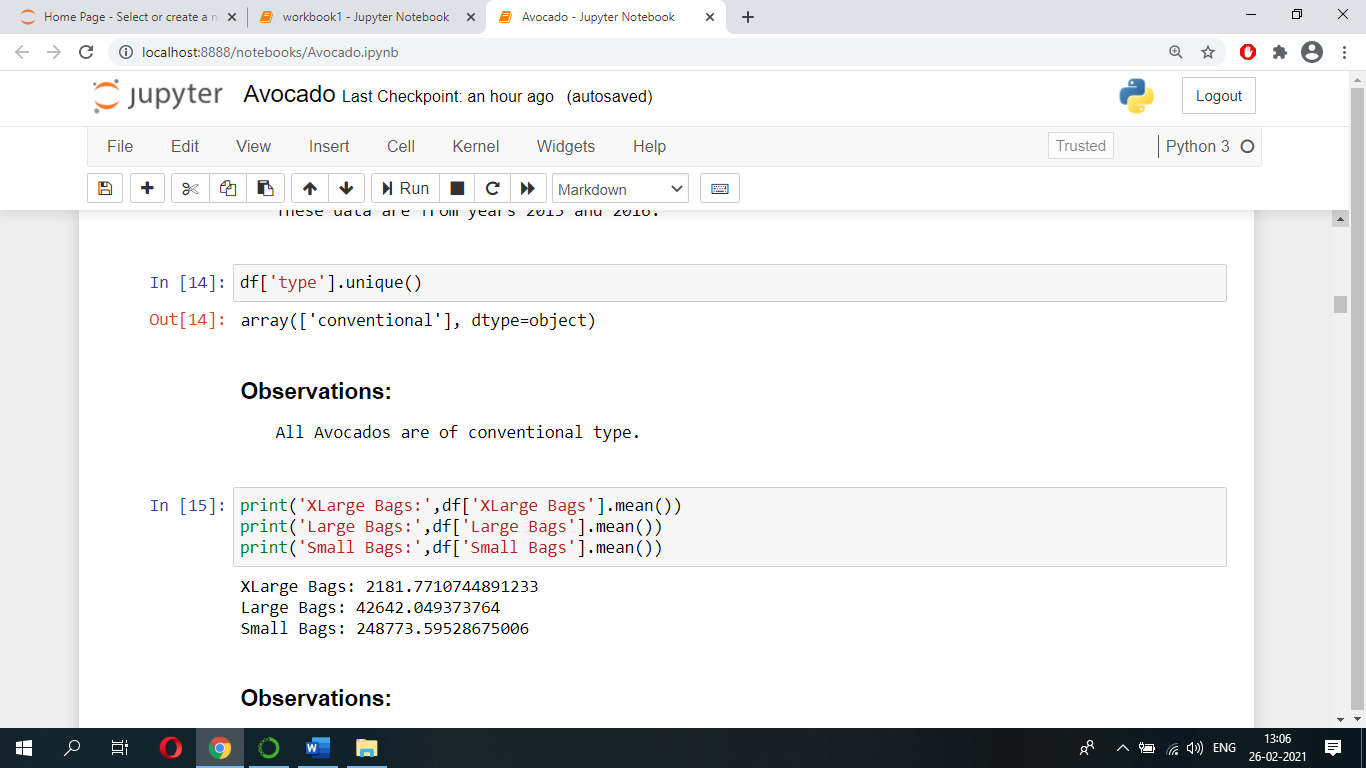
Let’s see we got any missing values in our dataset and what are the data-types of attributes by using info() function.

From above we can see that our dataset contains no missing values in it (we can also validate this using heatmap function) and attributes such as date, type, region is in ‘object’ data-types and rest of the attributes are in ‘float’ data-types.

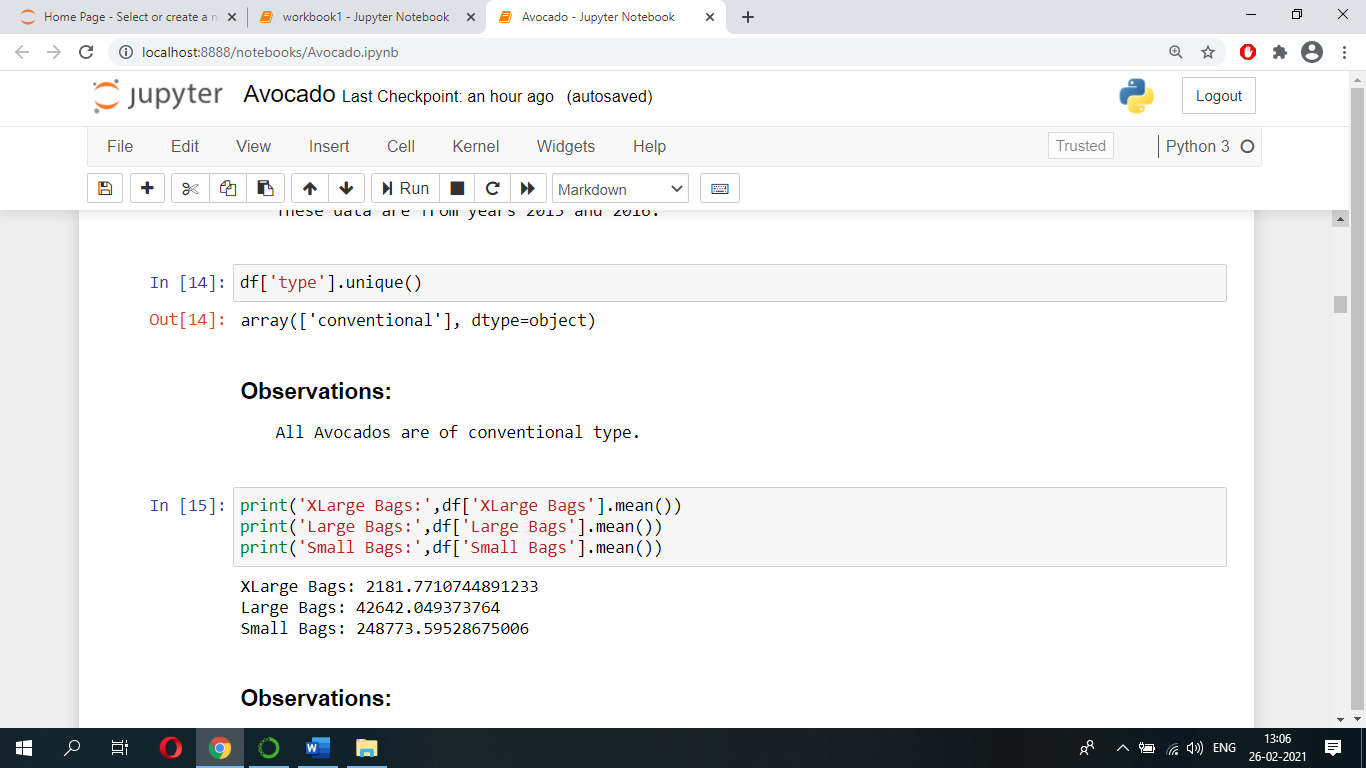
Let’s do some univariate analysis, first let’s find out what is the Average price of an avocado across cities/regions of United States using histogram. Below is the plot:

Let’s see what are the cities/regions of observations and the years of observations using unique() function.

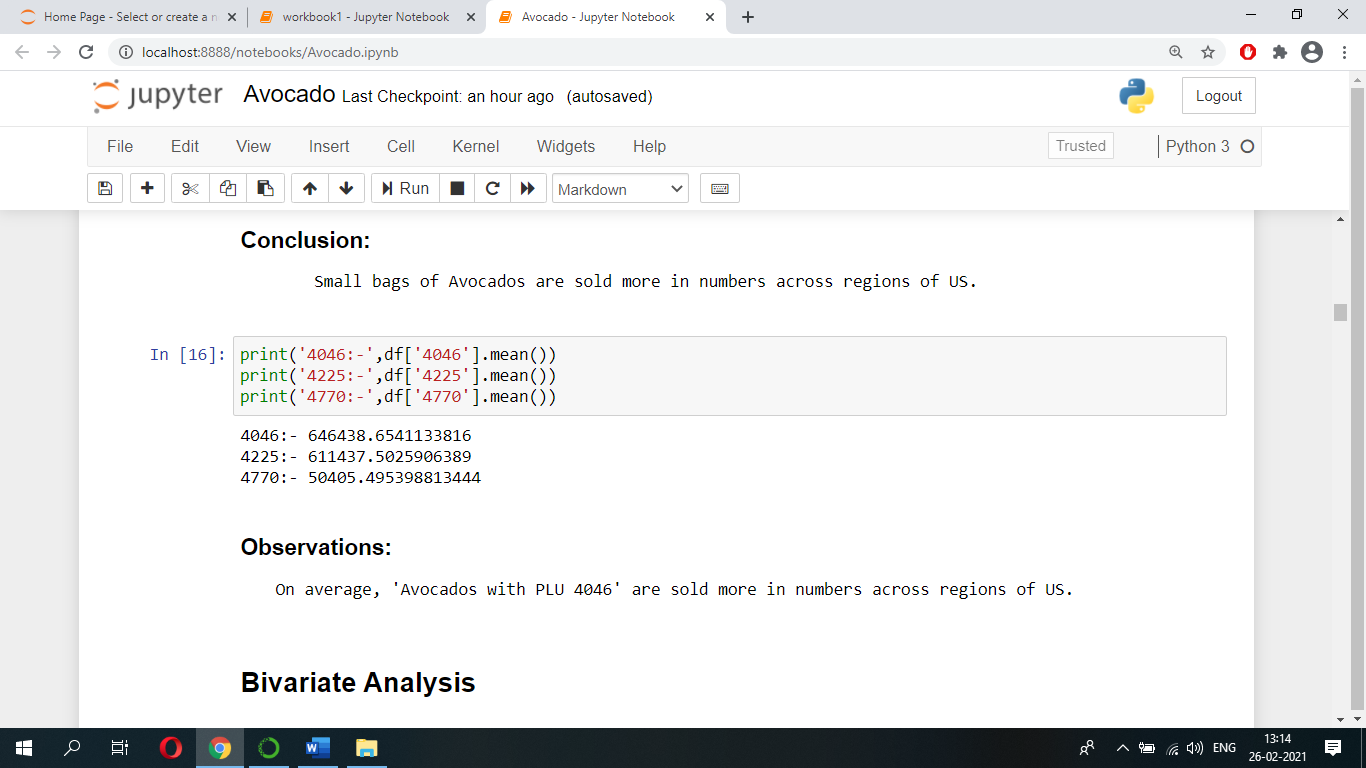
Let’s see what types of avocados are present in our dataset using unique() function.



Let’s see on average how many Small, Large, Xlarge bags of avocados are sold across cities/regions of US.

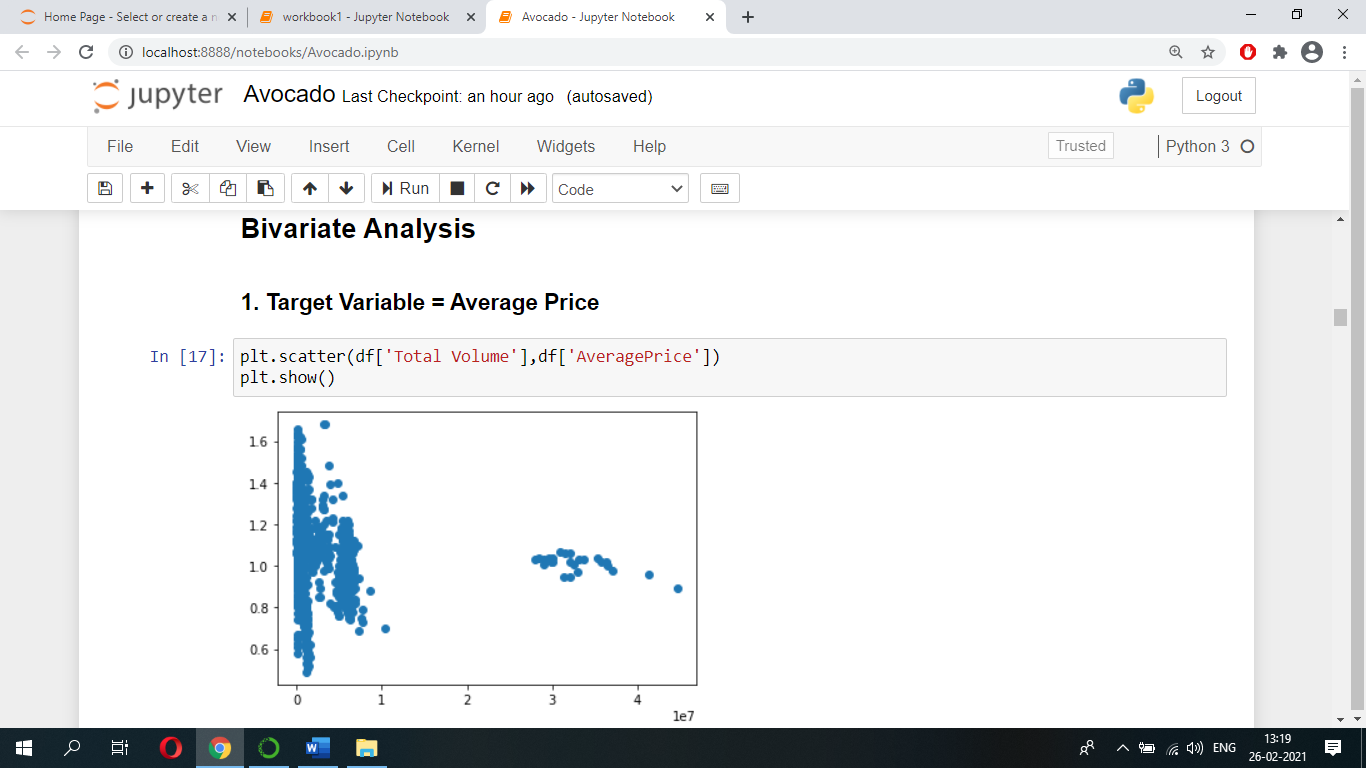


Let’s see on average which avocados i.e., ‘avocados with PLU 4046’, ‘avocados with PLU 4225, ‘avocados with PLU 4770’, are sold more across cities/regions of US.

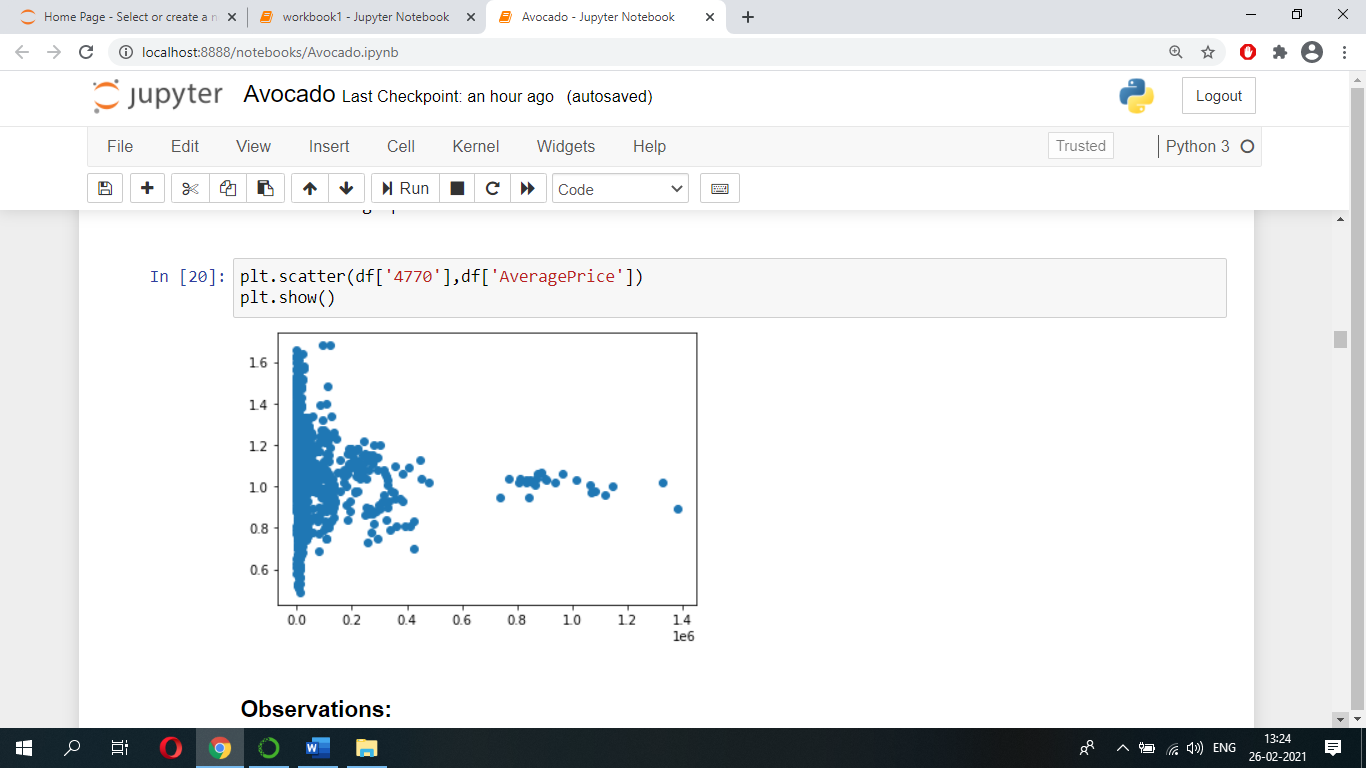
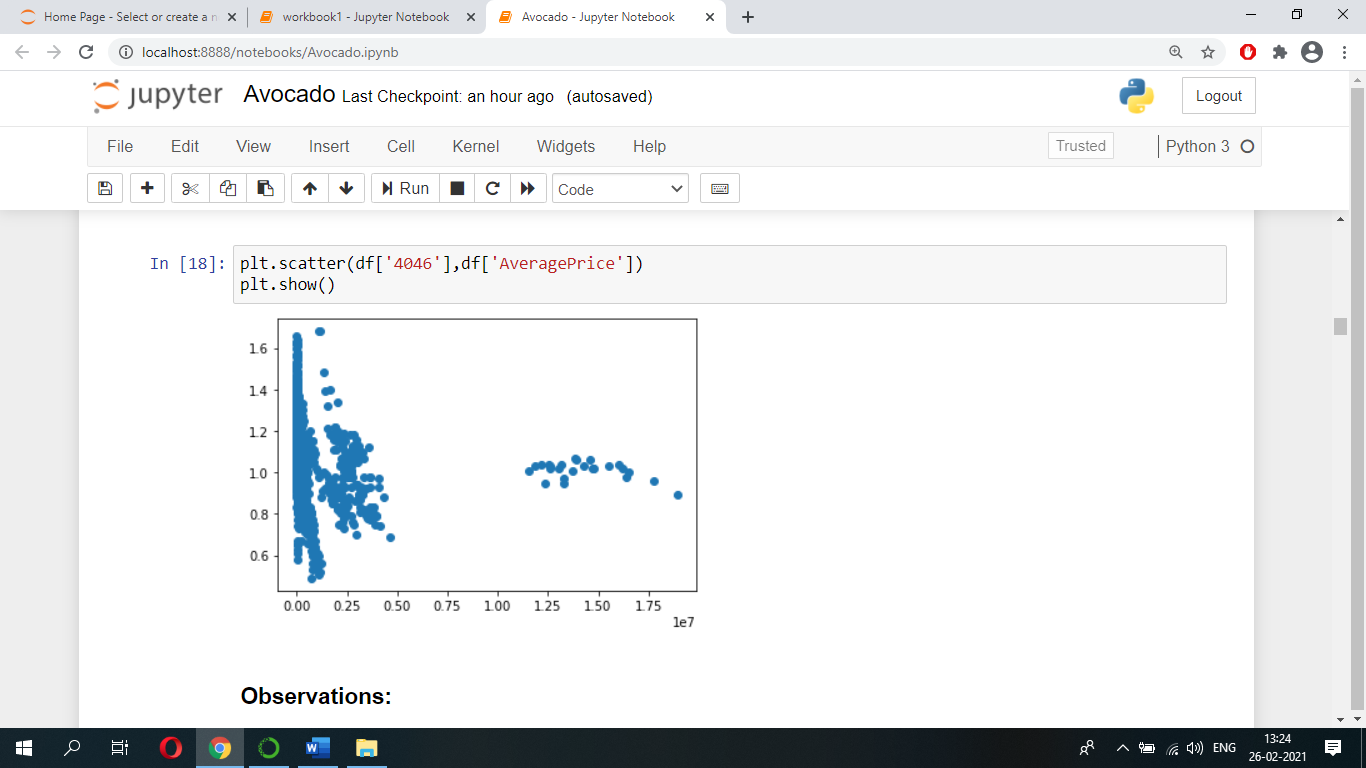
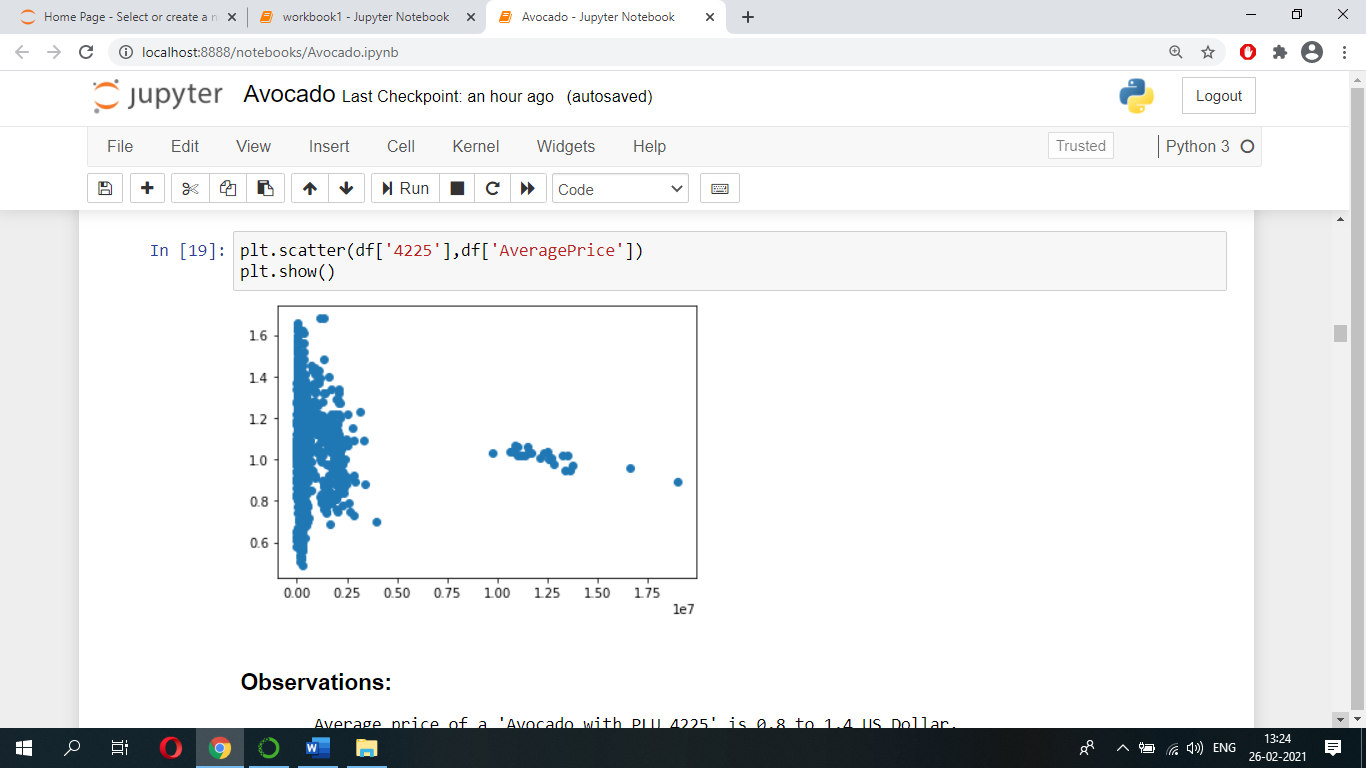


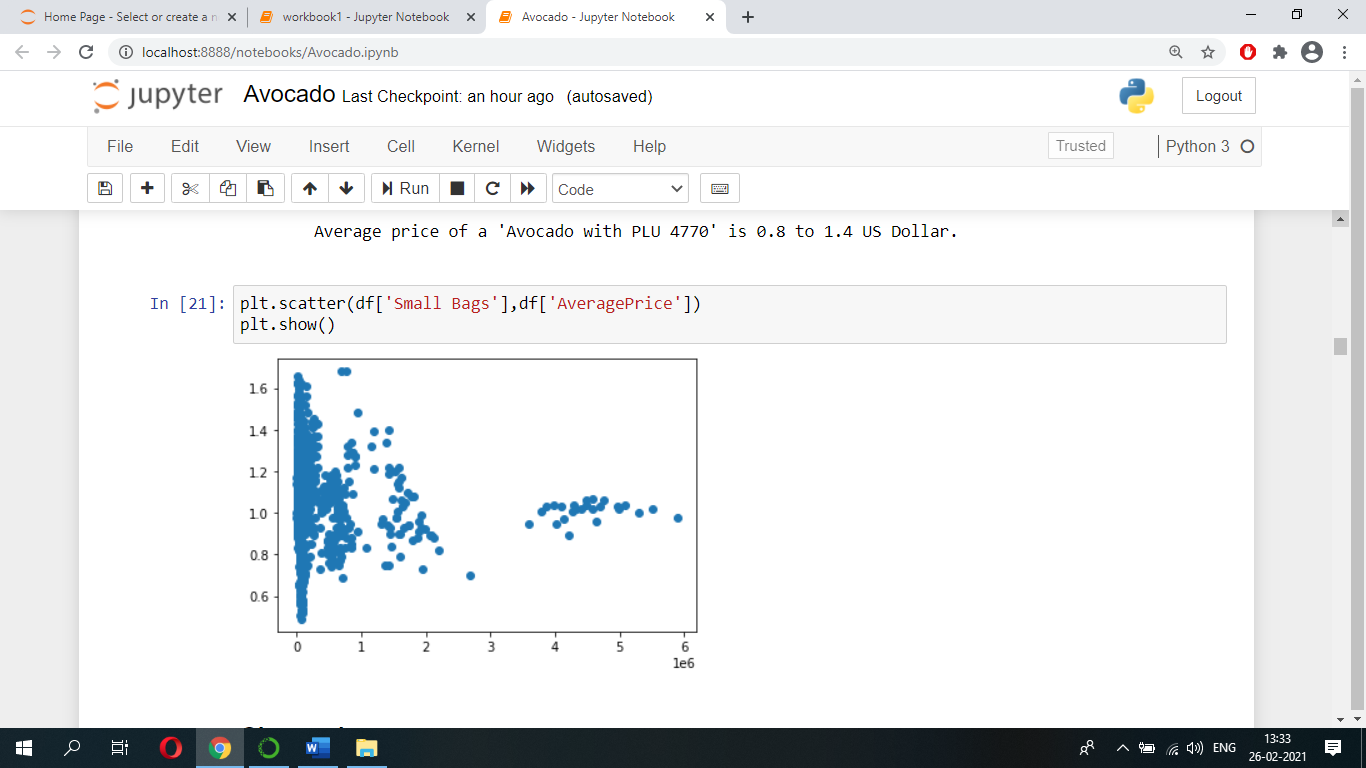
Let’s do some bivariate analysis because it will help us to visualize relations between different attributes and the target variable.

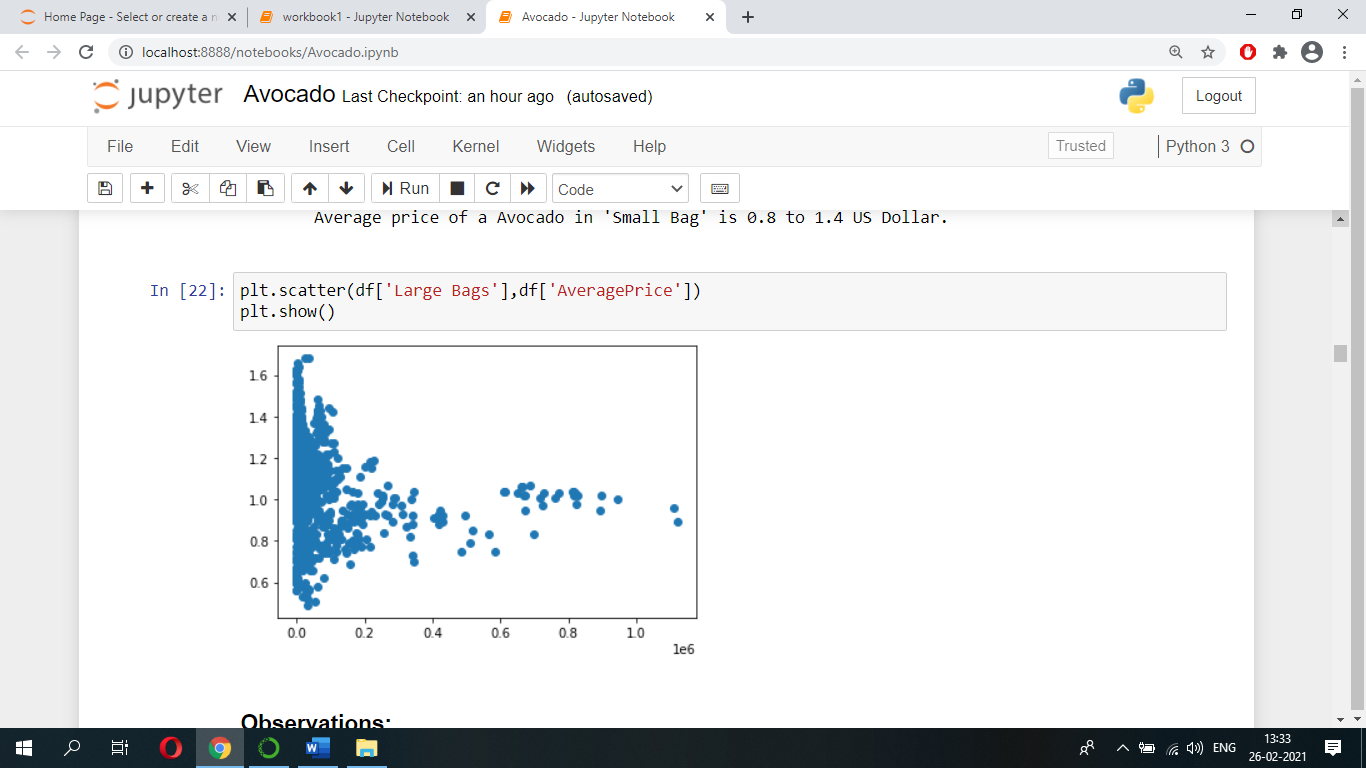
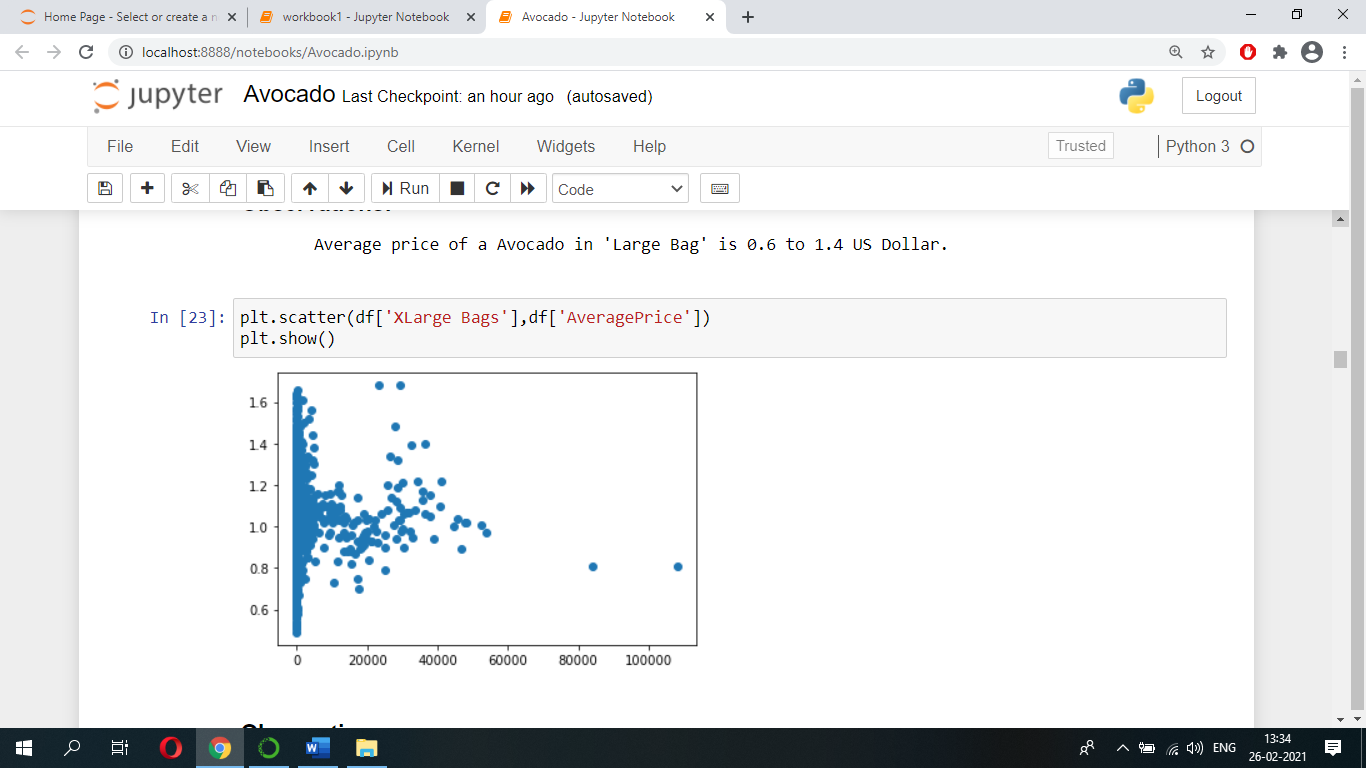
Let’s see how Average price varies when Total volume of avocados varies using scatterplot.



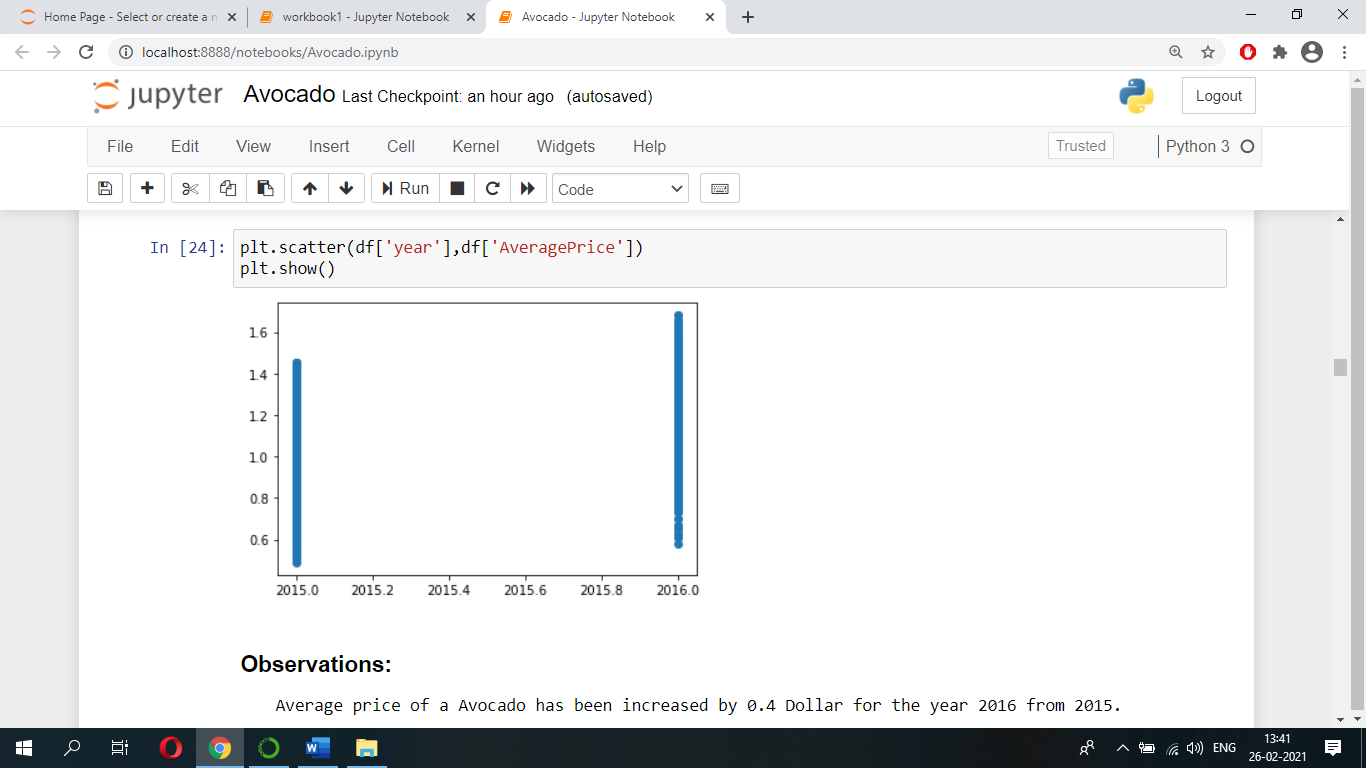
Let’s see what is the Average price of an ‘avocado with PLU 4046’, ‘avocado with PLU 4225, and ‘avocado with PLU 4770’ using scatterplot.



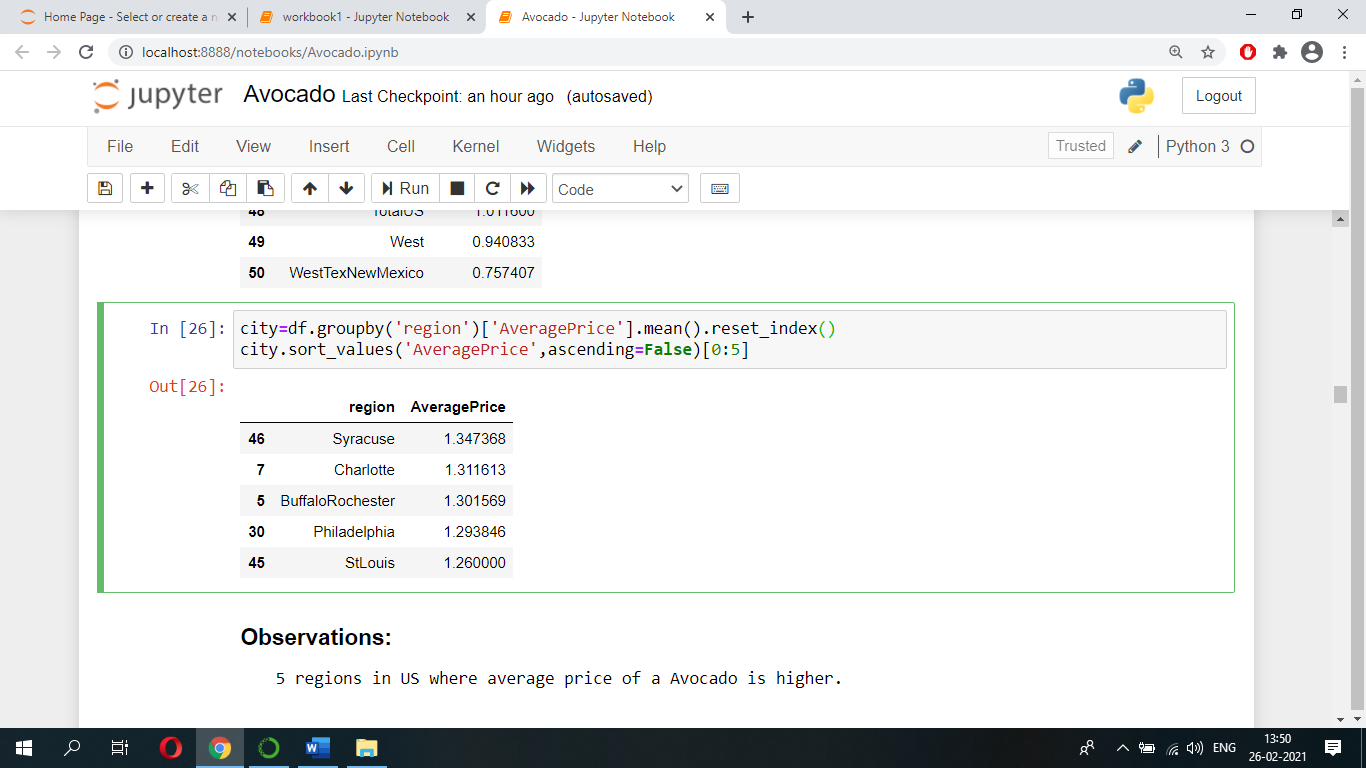
Let’s see what is the Average price of an avocado in Small, Large, Xlarge bags using scatterplot.

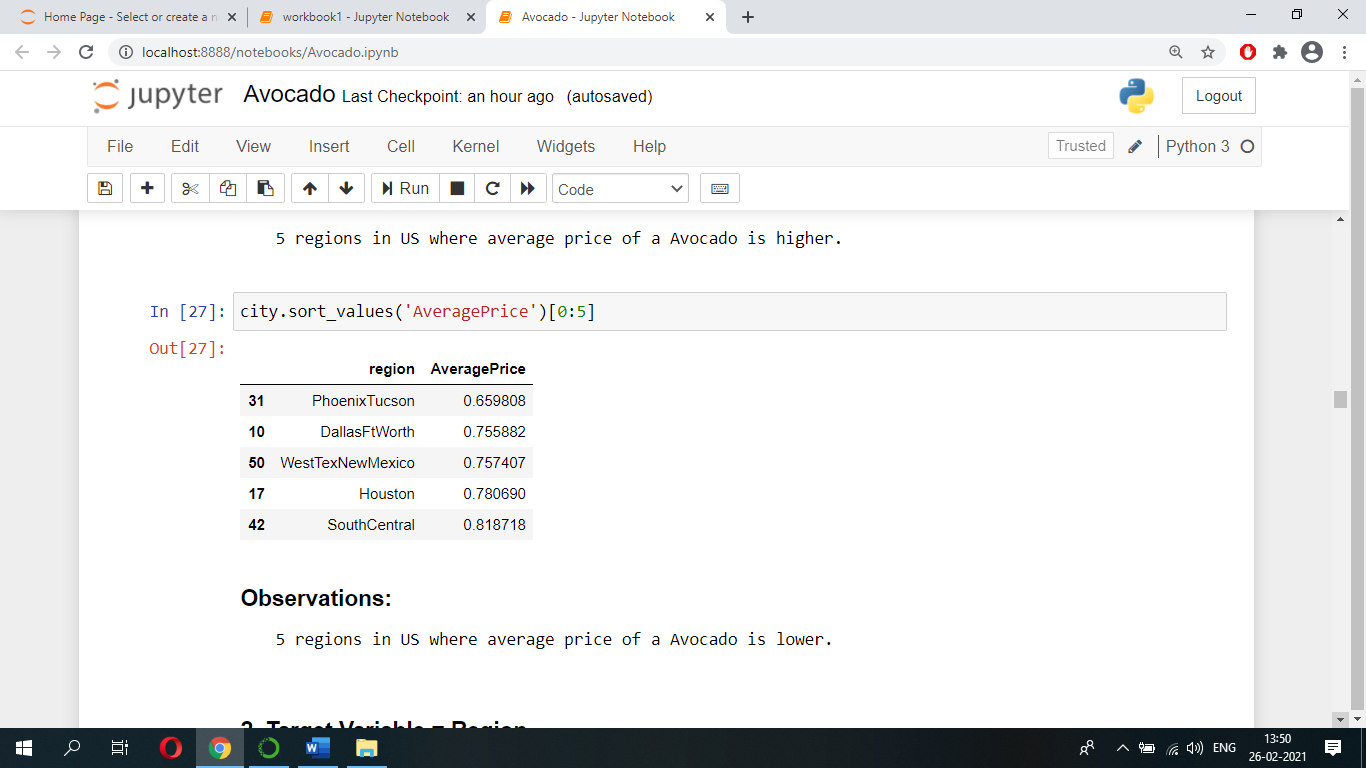


Let’s see whether avocado price has been increased for the year 2016 from 2015 using scatterplot.

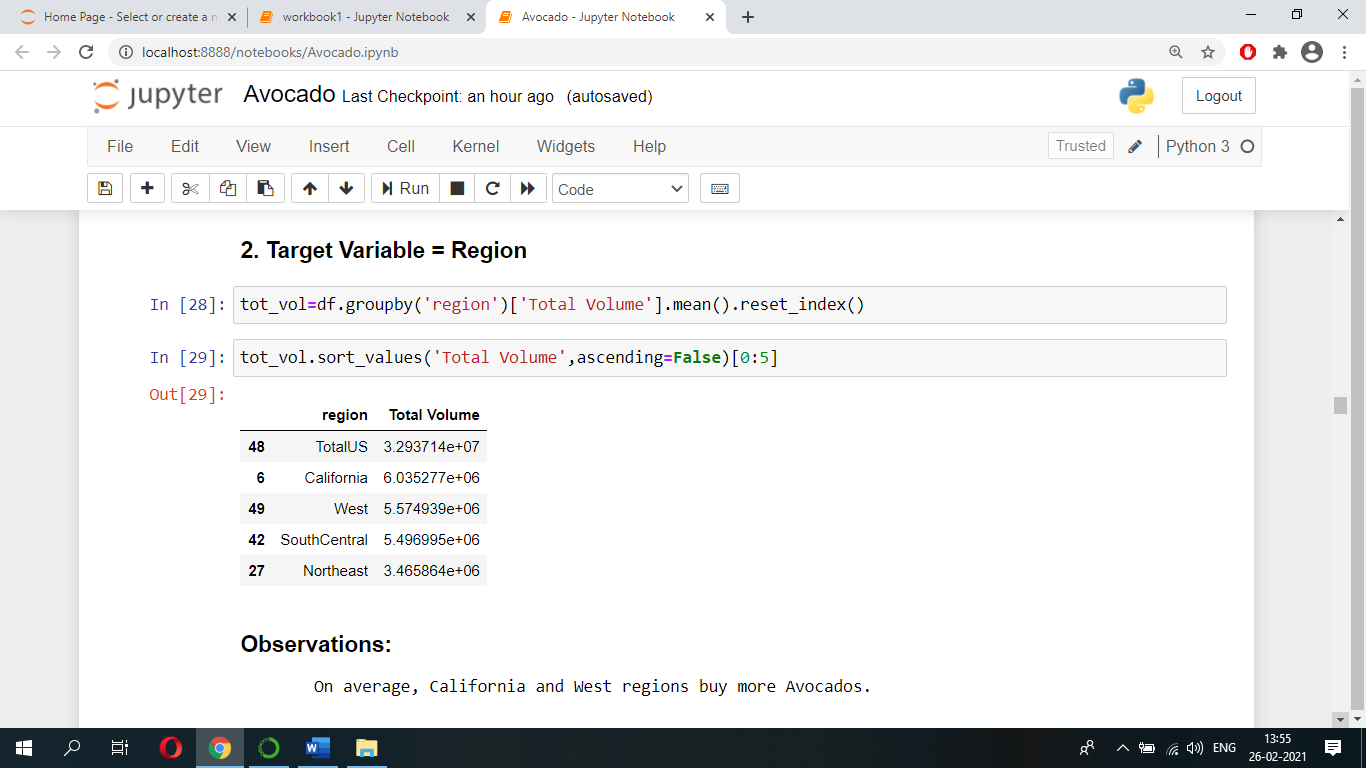


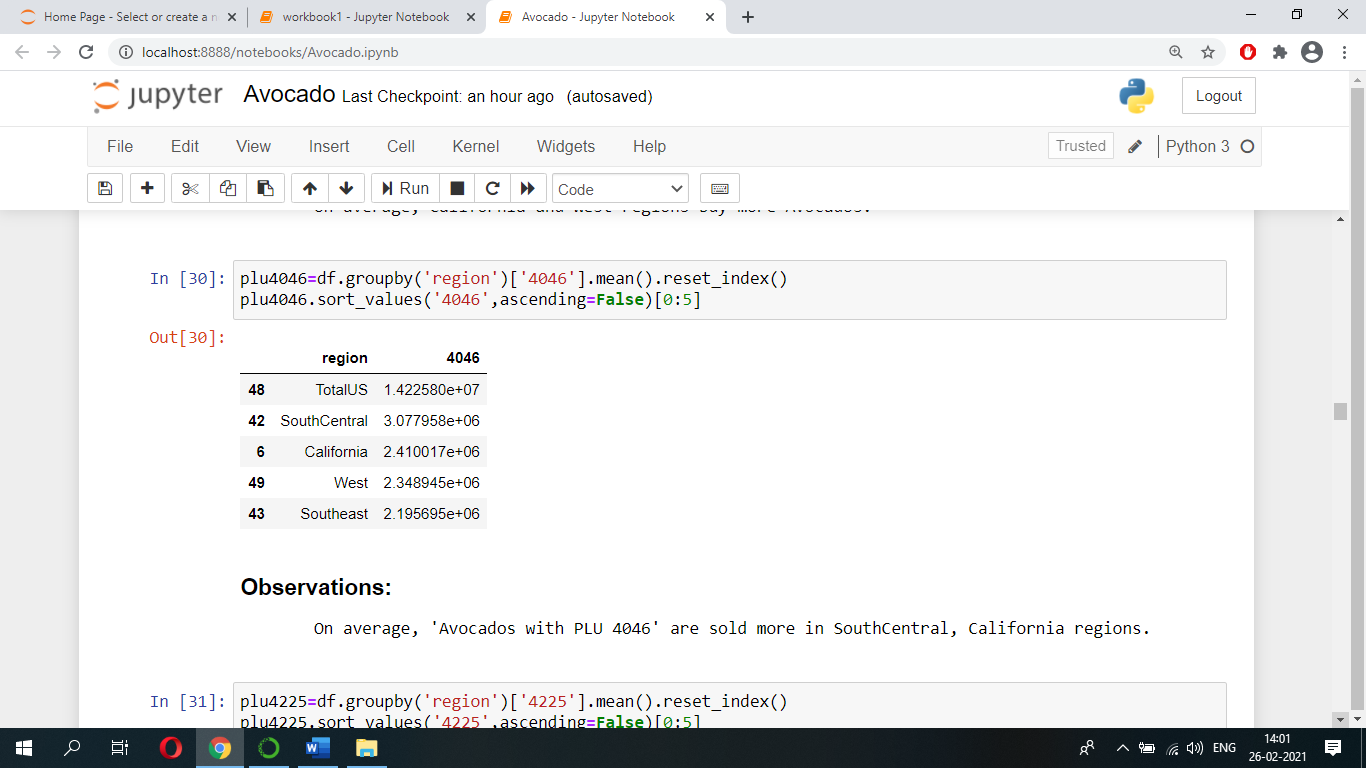
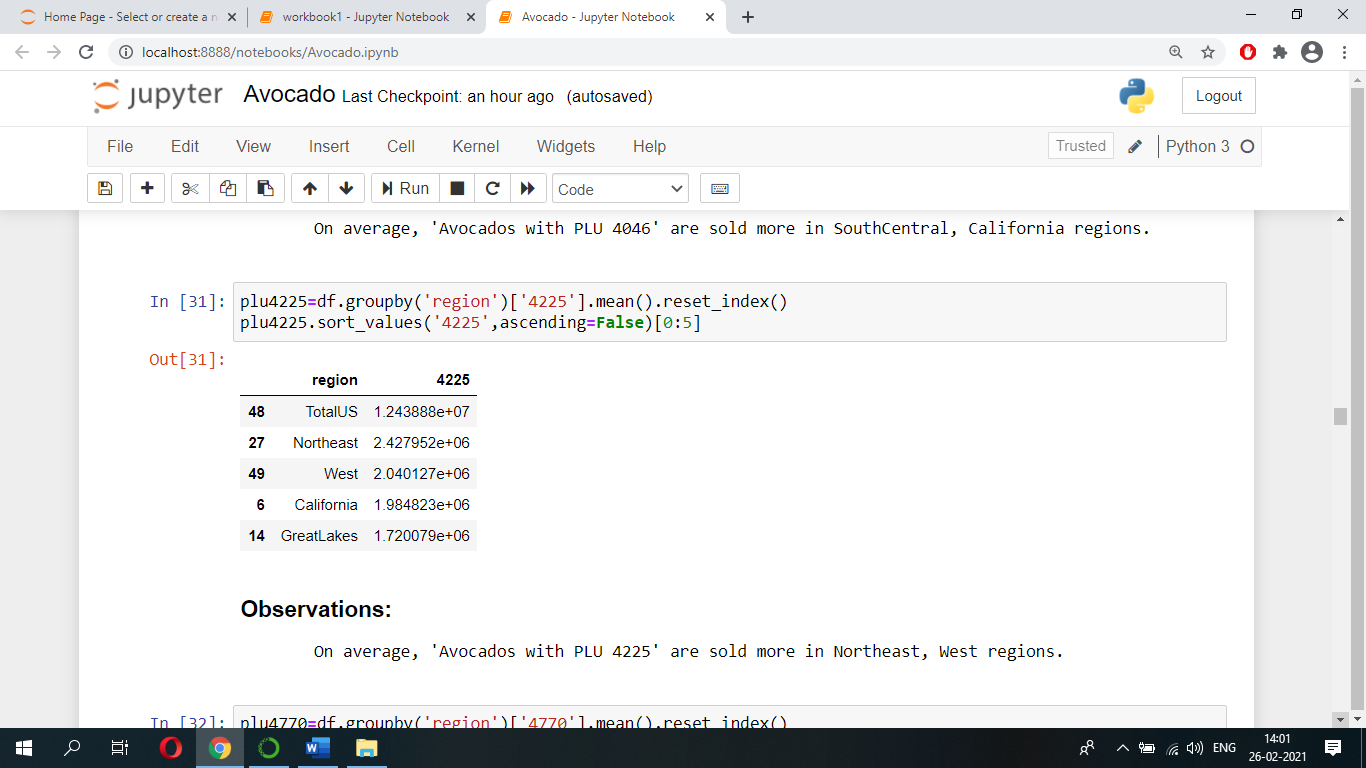
Let’s see on which cities/regions Average price of an avocado is higher when compared to others.

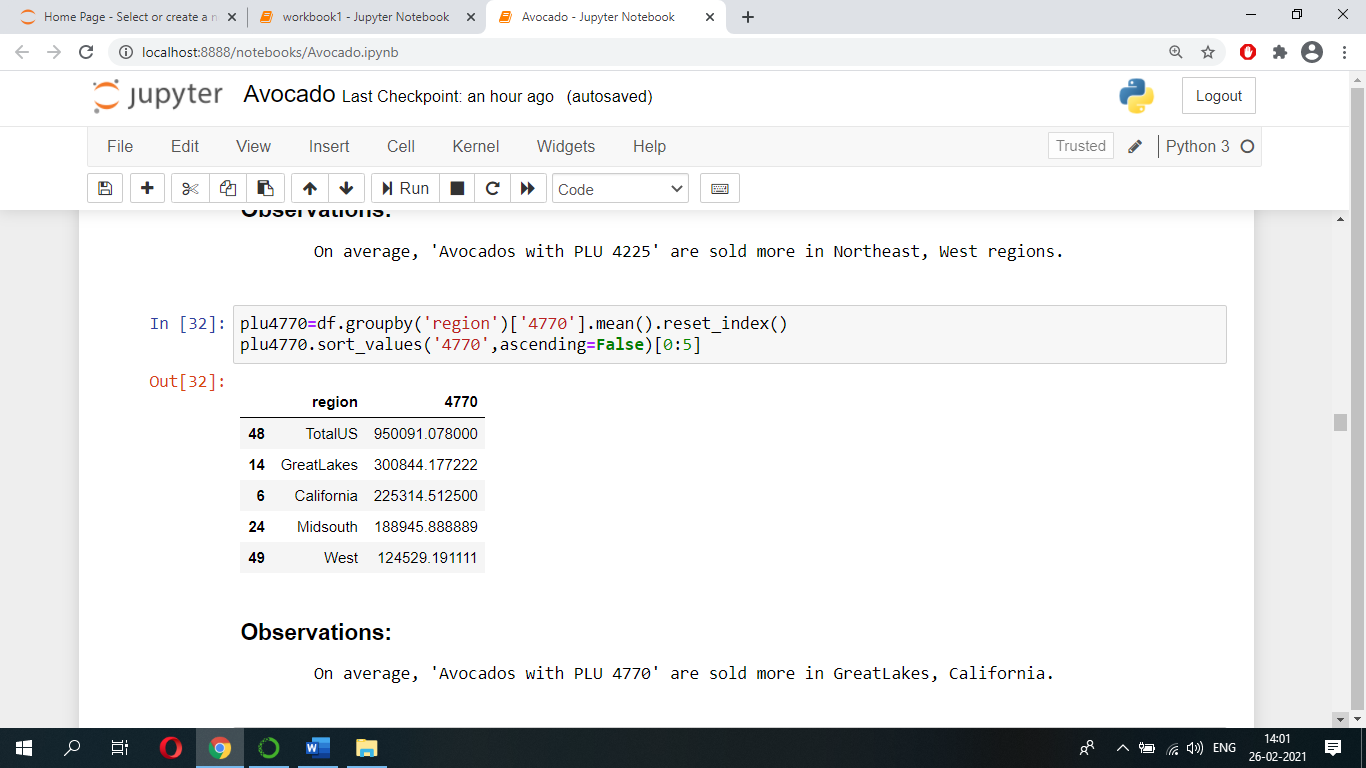


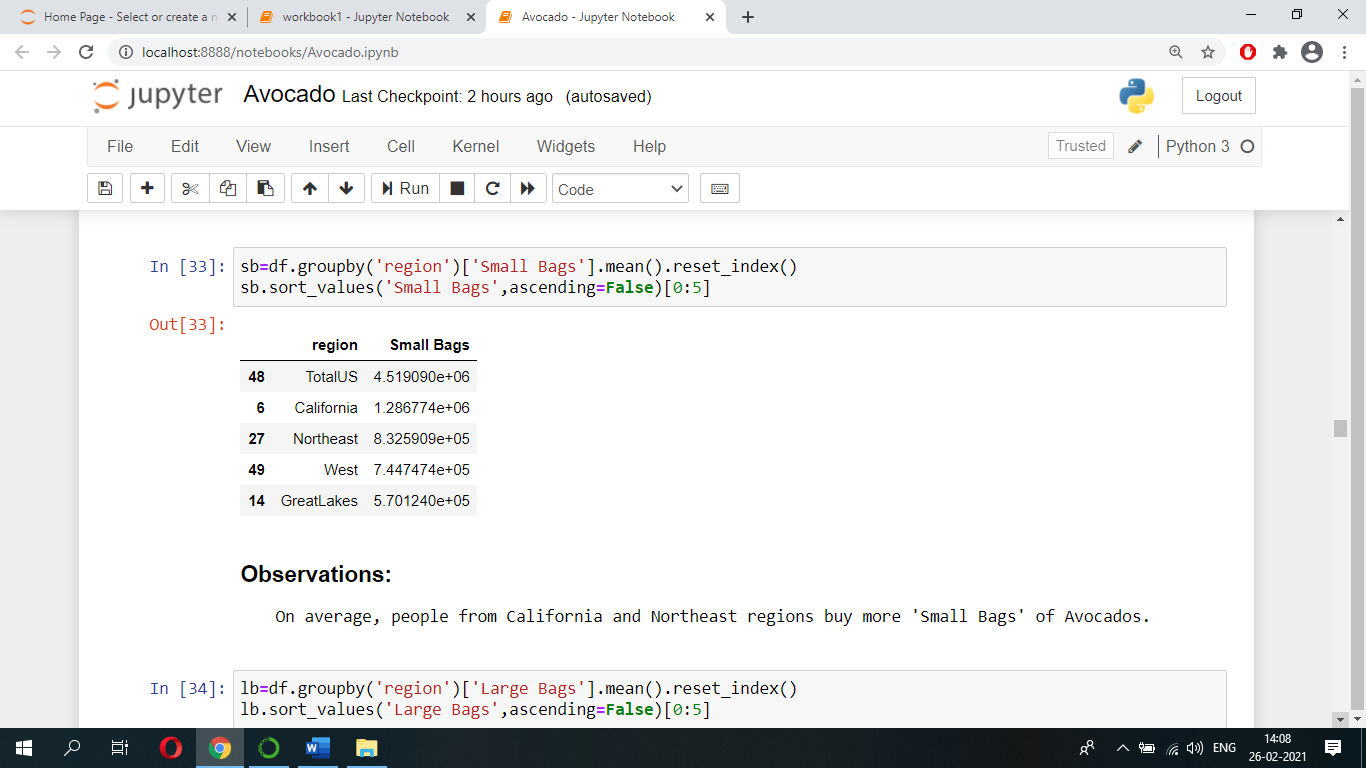
Let’s see on which cities/regions Average price of an avocado is lower when compared to others.

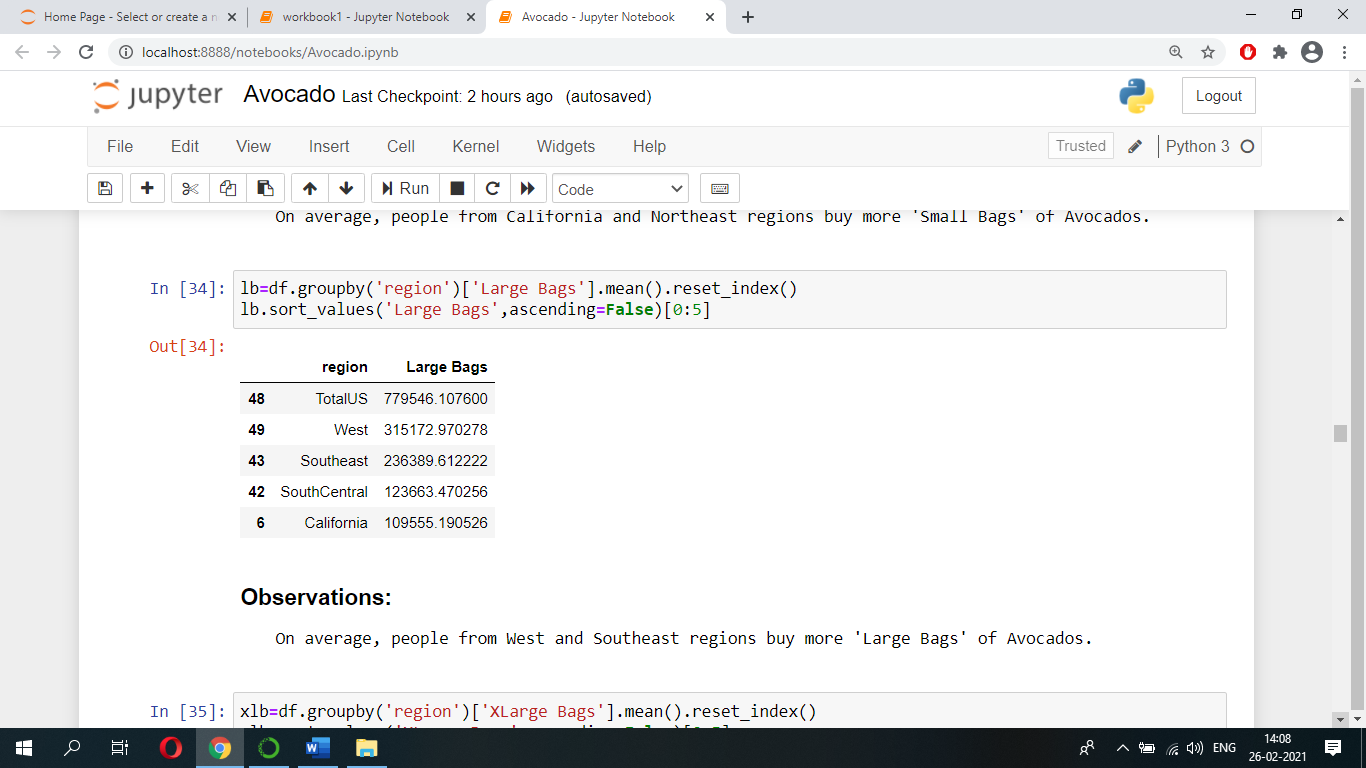
Let’s see on which cities/regions more volumes of avocados being sold.



Let’s see people from which cities/regions buy, ‘avocados with PLU 4046’, ‘avocados with PLU 4225, and ‘avocados with PLU 4770’, more in numbers.

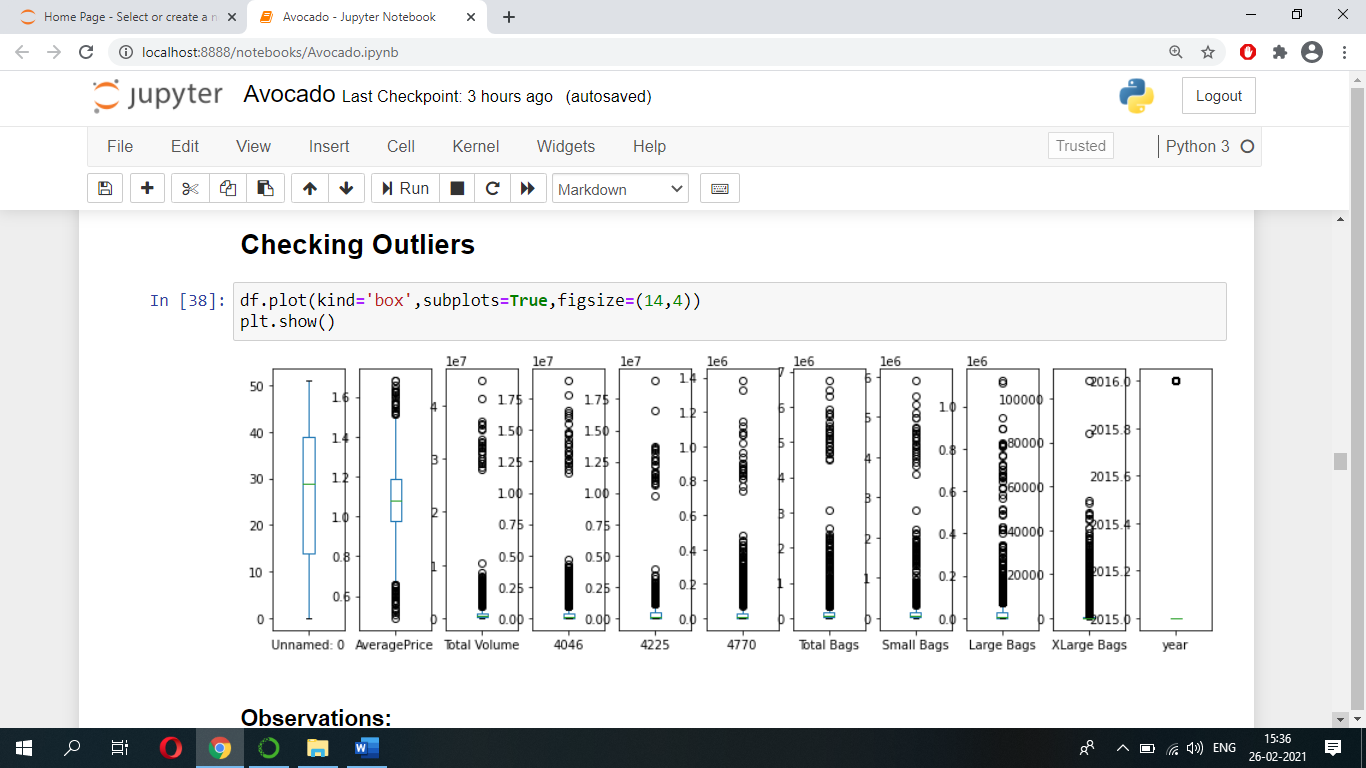


Let’s see people from which cities/regions buy, Small, Large, XLarge, bags of avocados, more in numbers.





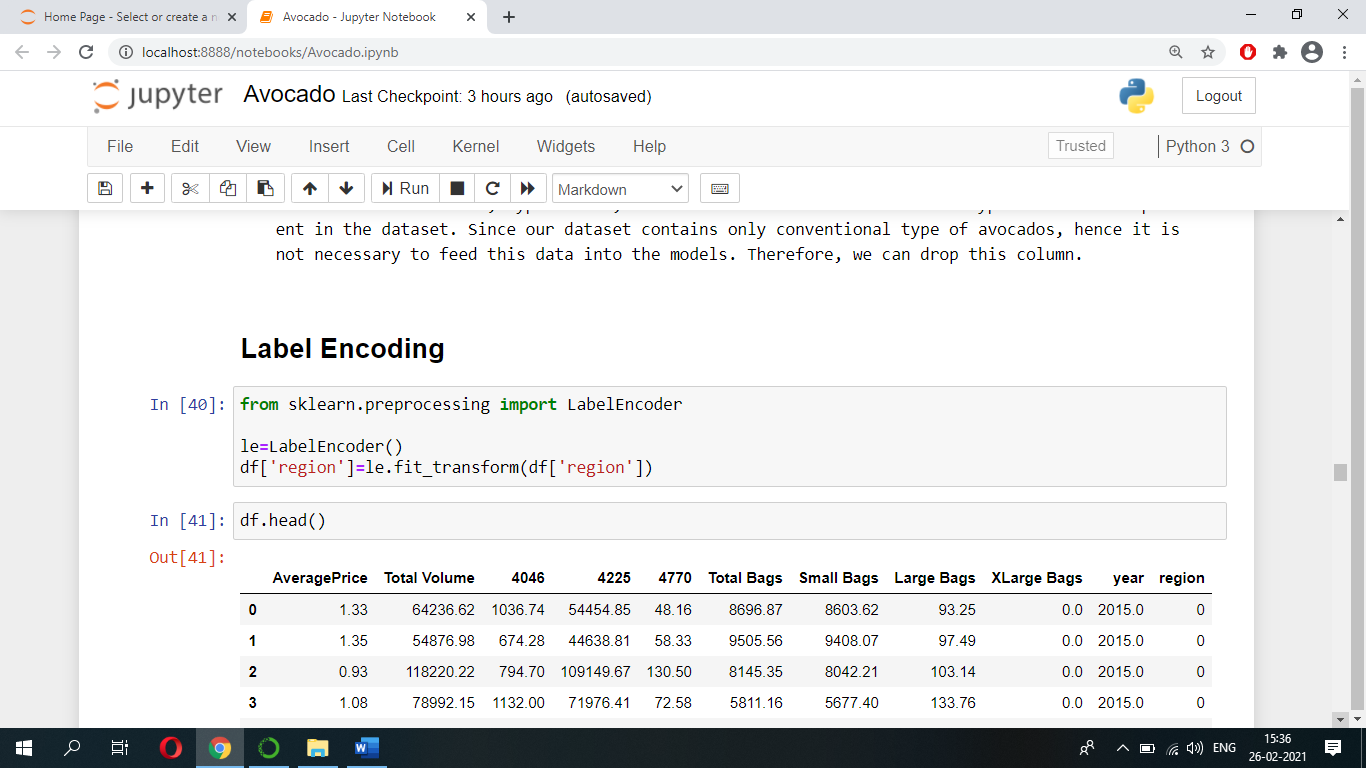
* **Summary Of EDA Findings**
* Findings of Univariate analysis are as follows:
* Average price of an avocado is 1 to 1.2 US Dollar across regions of United States.
* There are 51 cities/regions of observations.
* These data are from years of 2015 and 2016.
* Avocados are of two types; one is conventional type and another one is of organic type**.** But in our dataset all avocados are of conventional type.
* On average, 2181 XLarge Bags of avocados are sold across regions of United states.
* On average, 42642 Large Bags of avocados are sold across regions of United states.
* On average, 248773 Small Bags of avocados are sold across regions of United states. **Hence Small Bags of avocados are sold more in numbers across regions of United states.**
* **On average, ‘avocados with PLU 4046’ are sold more in numbers across regions of United States.**
* Findings of Bivariate analysis are as follows:
* As total volume of avocados increases Average price of an avocado decreases.
* Average price of an ‘avocado with PLU 4046’ is 0.6 to 1.2 US Dollar.
* Average price of an ‘avocado with PLU 4225’ is 0.8 to 1.4 US Dollar.
* Average price of an ‘avocado with PLU 4770 ‘is 0.8 to 1.4 US Dollar.
* Average price of an ‘avocado in a Small Bag is 0.8 to 1.4 US Dollar.
* Average price of an ‘avocado in a Large Bag is 0.6 to 1.4 US Dollar.
* Average price of an ‘avocado in a XLarge Bag is 0.8 to 1.4 US Dollar.
* Average price of an avocado has been increased by 0.4 US Dollar for the year 2016 from 2015.
* Syracuse, Charlotte, Buffalo Rochester, Philadelphia, StLouis are the cities/regions where Average price of an avocado is higher.
* Phoenix Tucson, Dallas Ft Worth, WestTexNewMexico, Houston, South Central are the cities/regions where Average price of an avocado is lower.
* On average, people from California, West regions of US buy more avocados.
* On average, ‘avocado with PLU 4046’ are sold more in South Central, California regions.
* On average, ‘avocado with PLU 4225 are sold more in North East and West regions.
* On average, ‘avocado with PLU 4770 are sold more in Great Lakes and, California regions.
* On average, people from California and North East regions buy more Small Bags of avocados.
* On average, people from West and South East regions buy more Large Bags of avocados.
* On average, people from California and Great Lakes regions buy more XLarge Bags of avocados.
* **Data Pre-processing**
* **Checking Outliers**

Let’s check whether the dataset has any outliers by plotting boxplots. Below is the plot:

From above boxplots it is evident that dataset contains outliers.

* **Label Encoding**

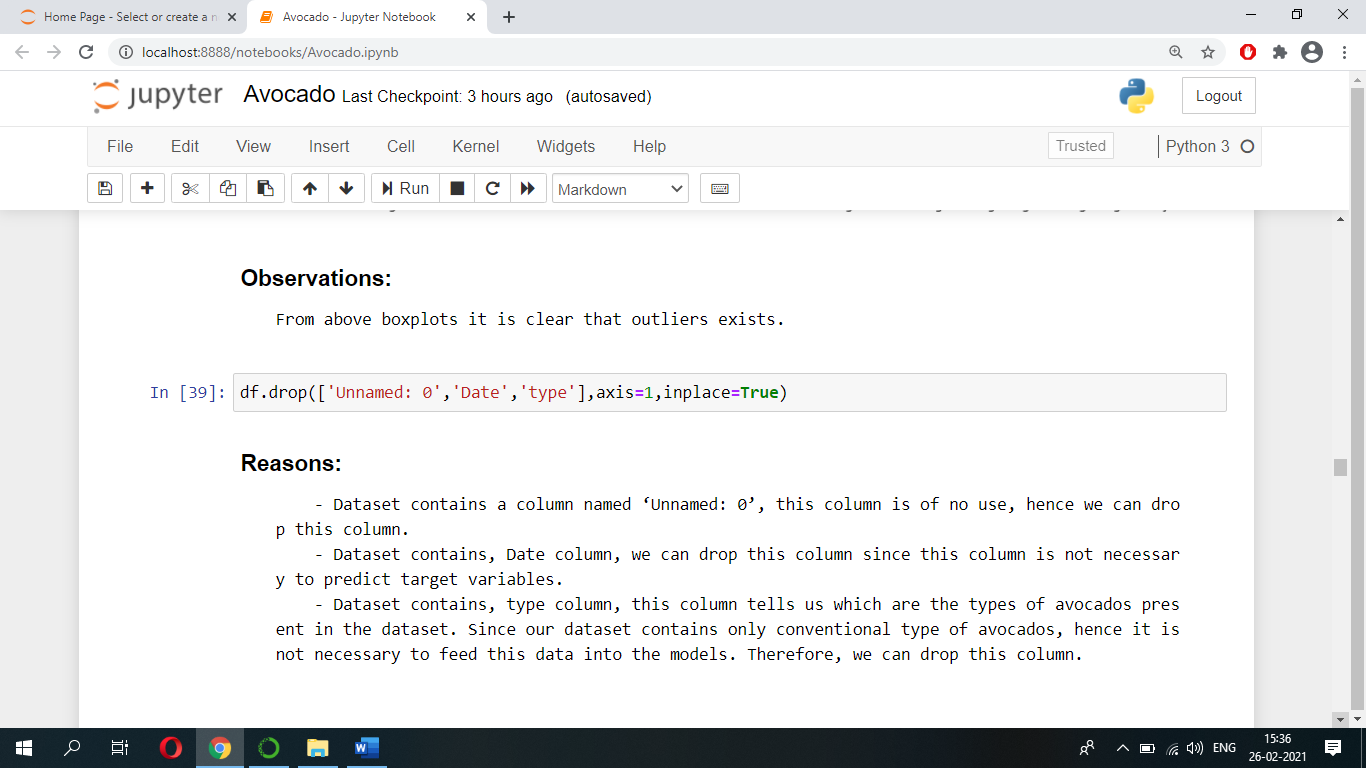
Before removing outliers, we should convert all columns values into numerical values. Since region column is in ‘object’ data-type i.e., its values are of categorical type, hence we should convert these into numerical values. Therefore, we are using Label Encoder to convert these categorical values into numerical values.

****

* **Dropping Columns:**

Dataset contains a column named ‘Unnamed: 0’, this column is of no use, hence we can drop this column.

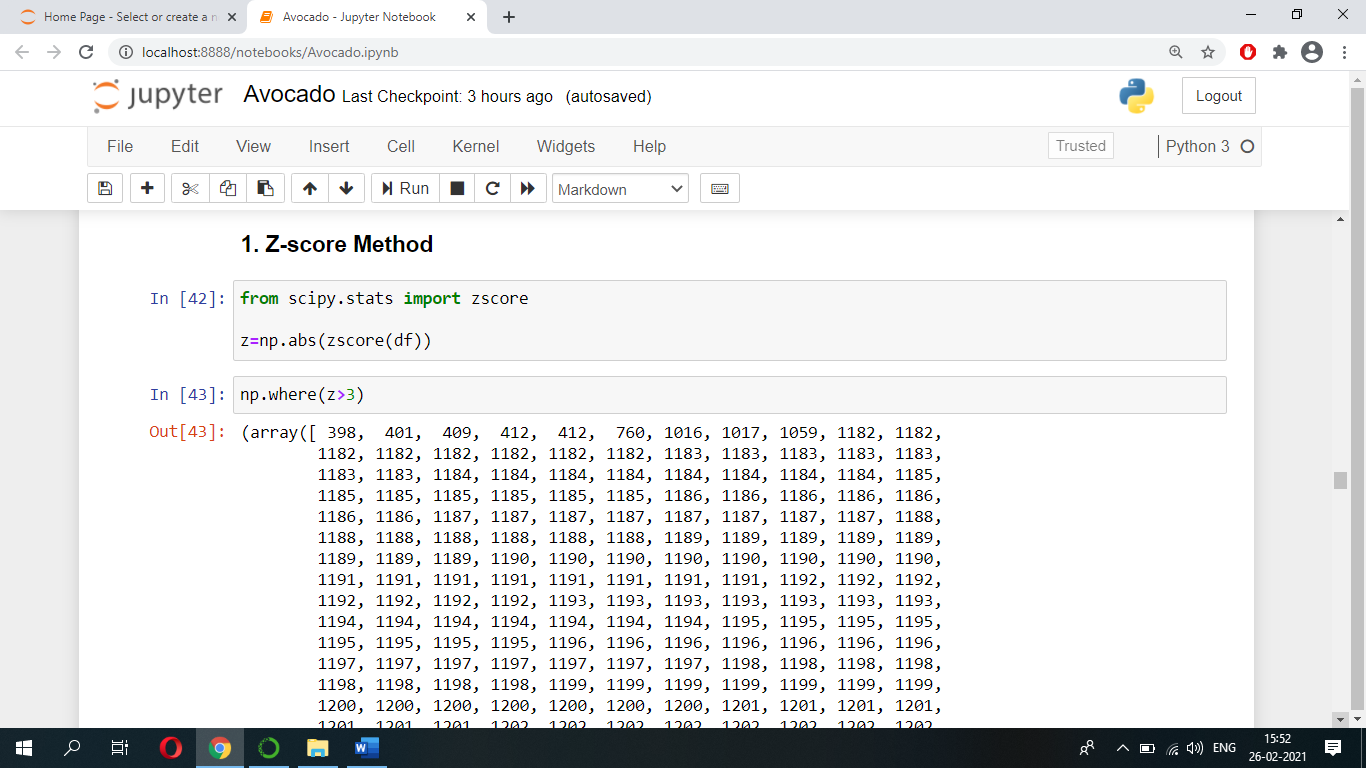
Dataset contains, Date column, we can drop this column because this column is not necessary to predict target variables.

Dataset contains, type column, this column tells us which type of avocados are present in our dataset. Since our dataset contains only conventional type of avocados, therefore it is not necessary to feed this data into the models. Hence, we can drop this column.

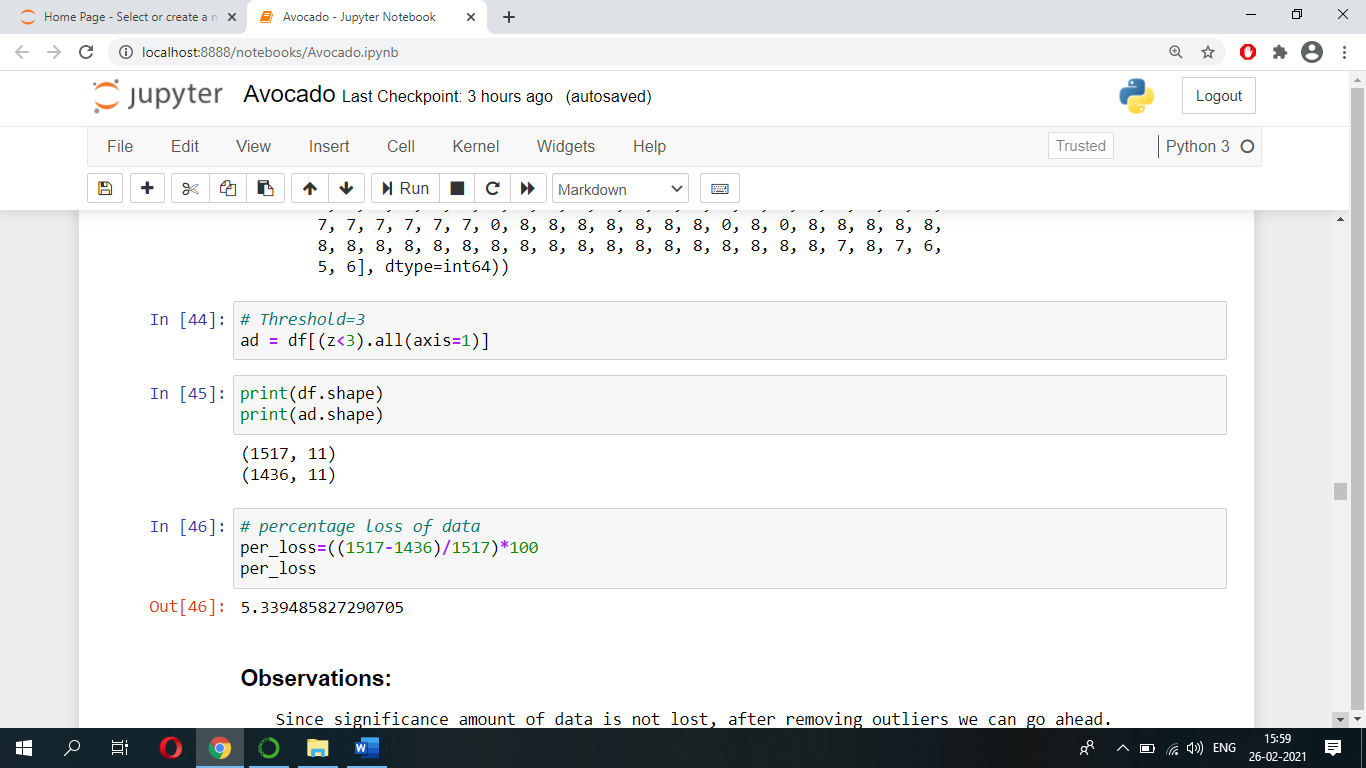
* **Removing Outliers**

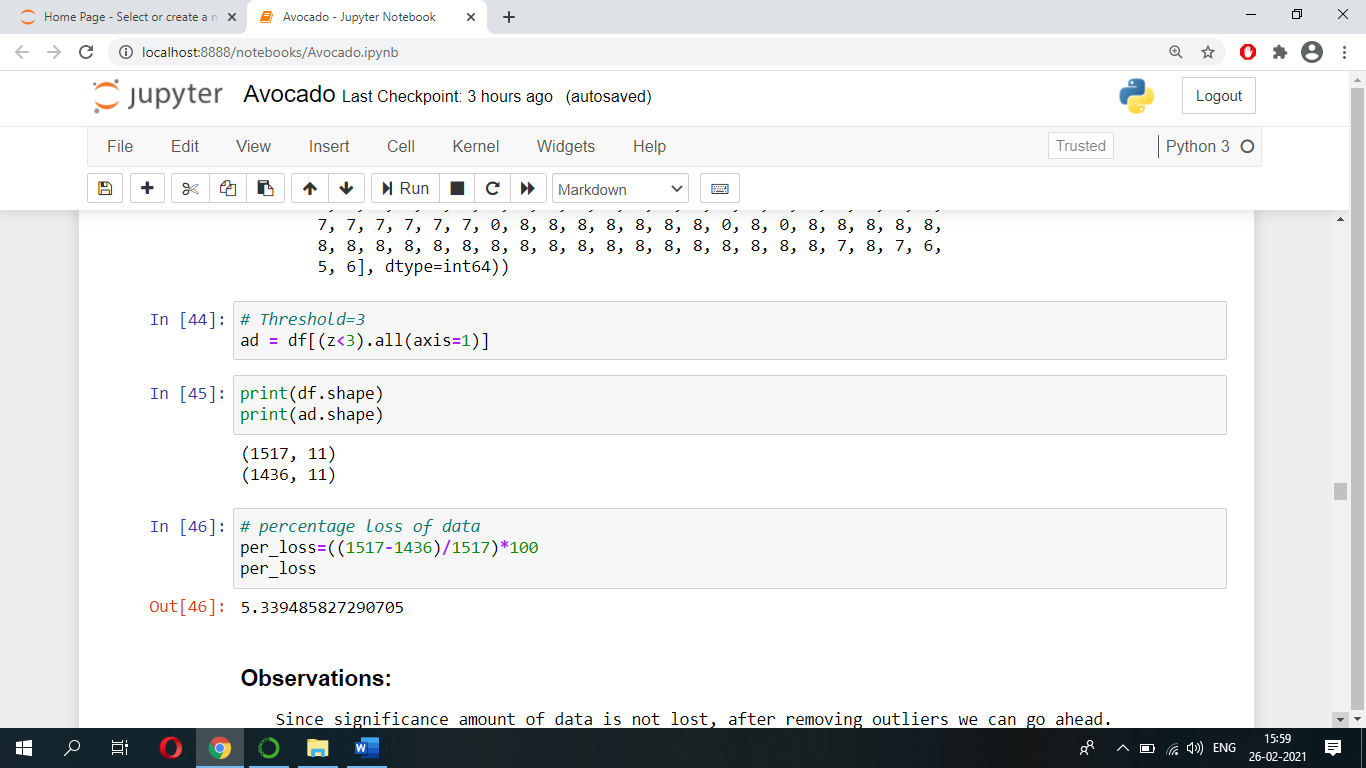
We have several methods to remove outliers, but the two widely used methods to remove outliers are, Z-score method and IQR method. Let’s select Z-score method to remove outliers.

Let’s import ‘zscore’ function from ‘scipy.stats’ and find out the zscore value of columns values.



Let’s consider, any value beyond +/-3 Standard Deviation i.e., let’s take Threshold=3, as outliers. And if any such values exist, then let’s remove it. After removing outliers save the data into ‘ad’.

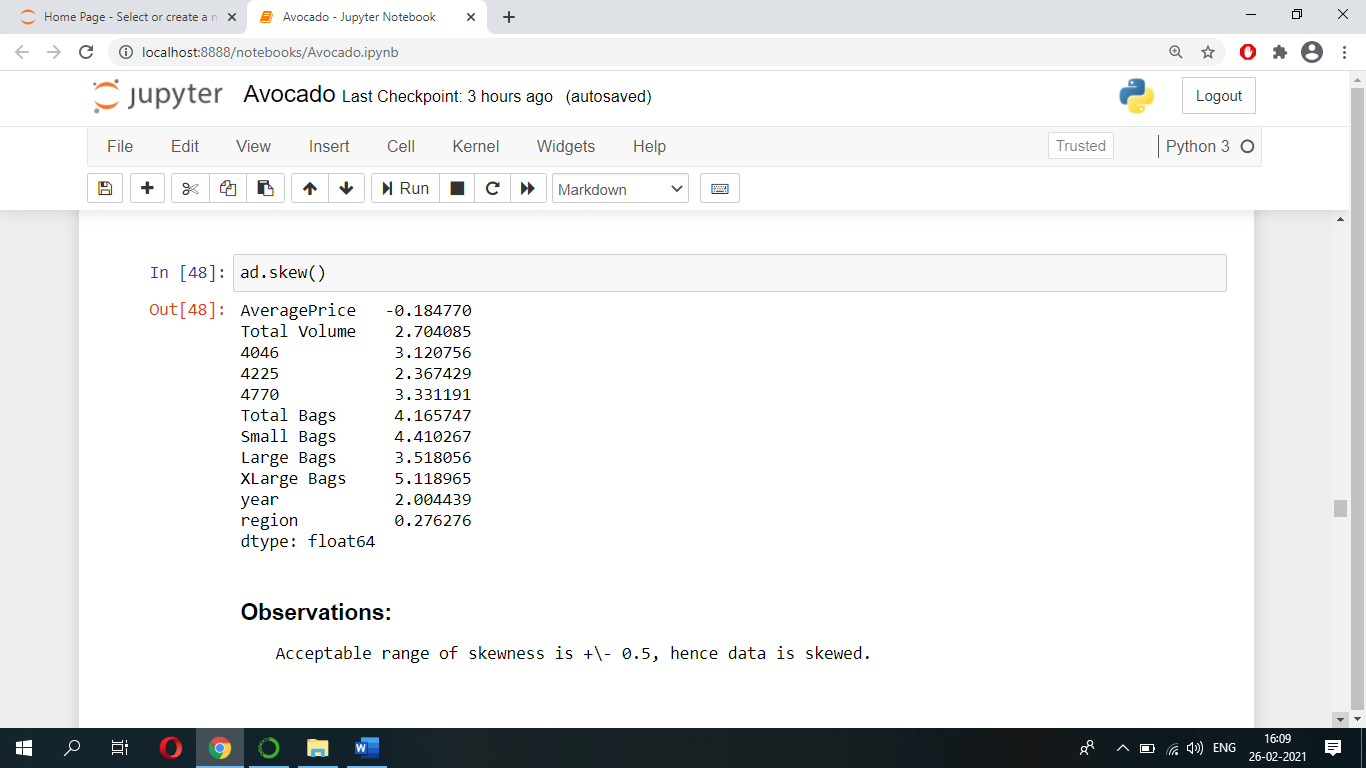


Now let’s check what is the percentage loss of data.

Since significant amount of data is not lost after removing outliers we can go ahead.

* **Checking Skewness**

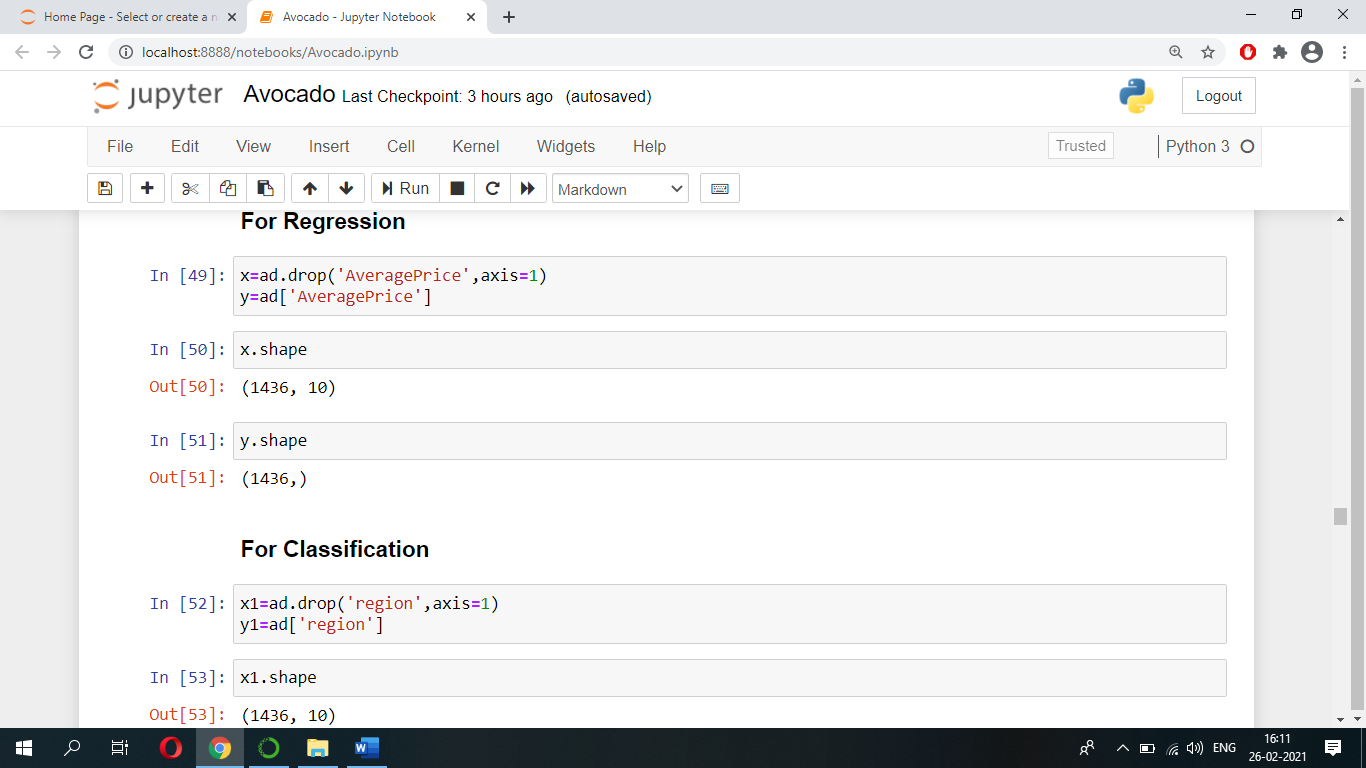
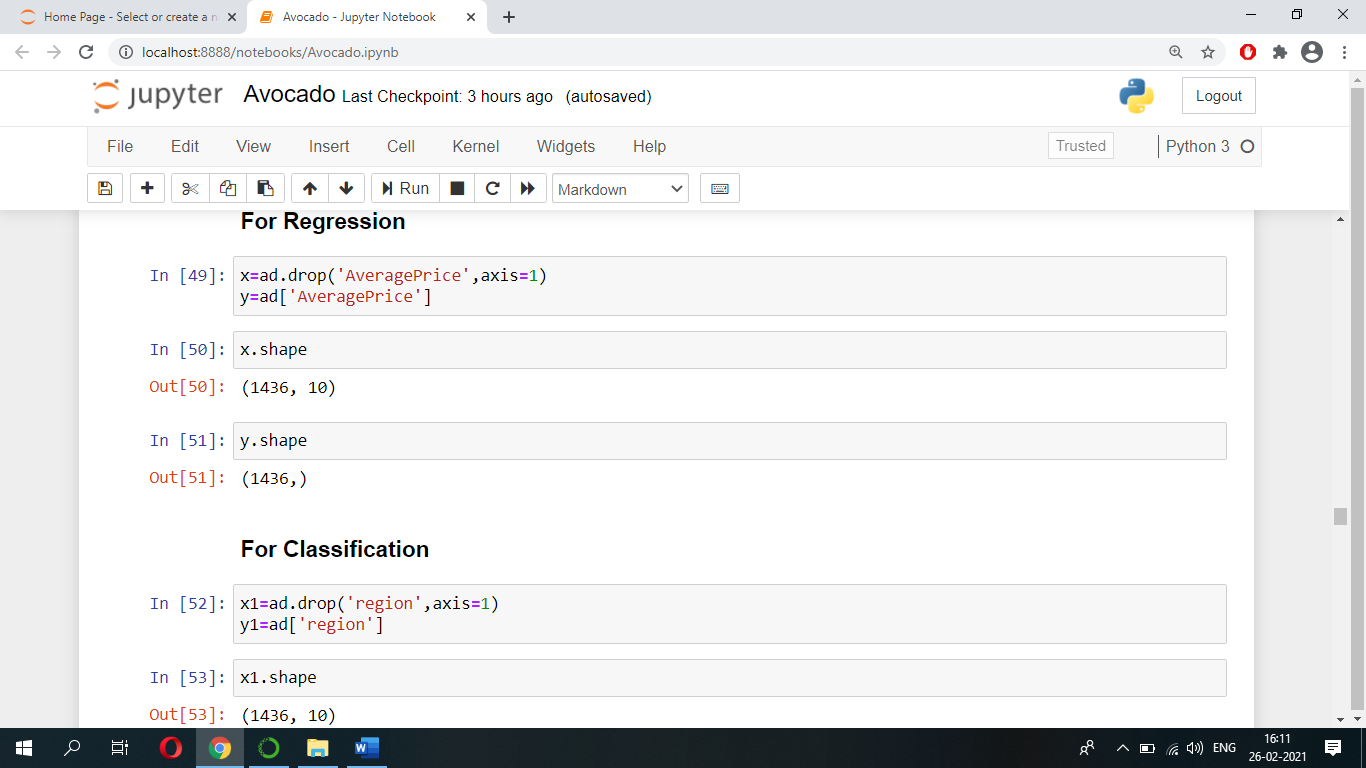
We can check whether data is skewed or not using histograms or we can use skew() function to check the same.



Acceptable range of skewness is +/- 0.5, hence from above it is clear that our data is skewed.

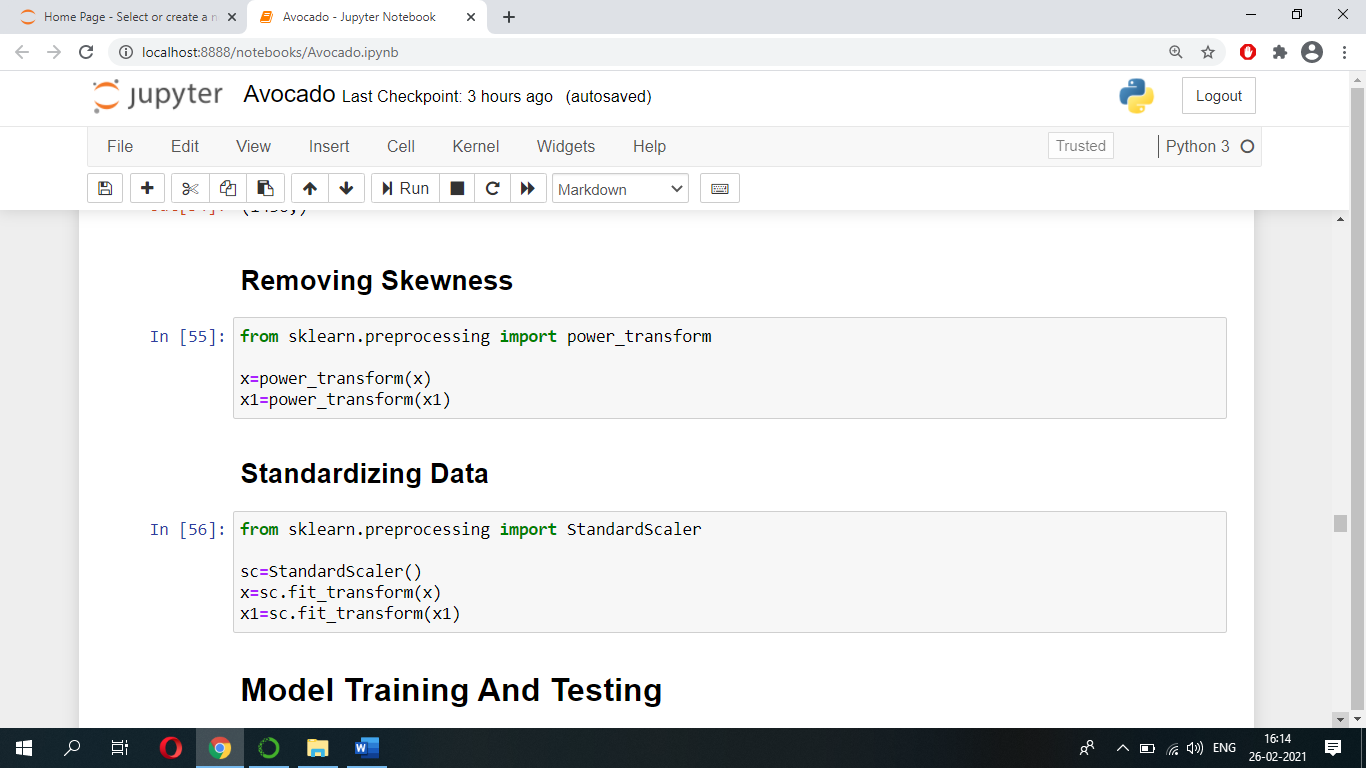
* **Splitting Data**

Let’s separate the label and features i.e., in order to predict ‘AveragePrice’ (Regression Problem) separate ‘AveragePrice’ from rest of the features and in order to predict ‘region’ (Classification Problem) separate ‘region’ from rest of the features.



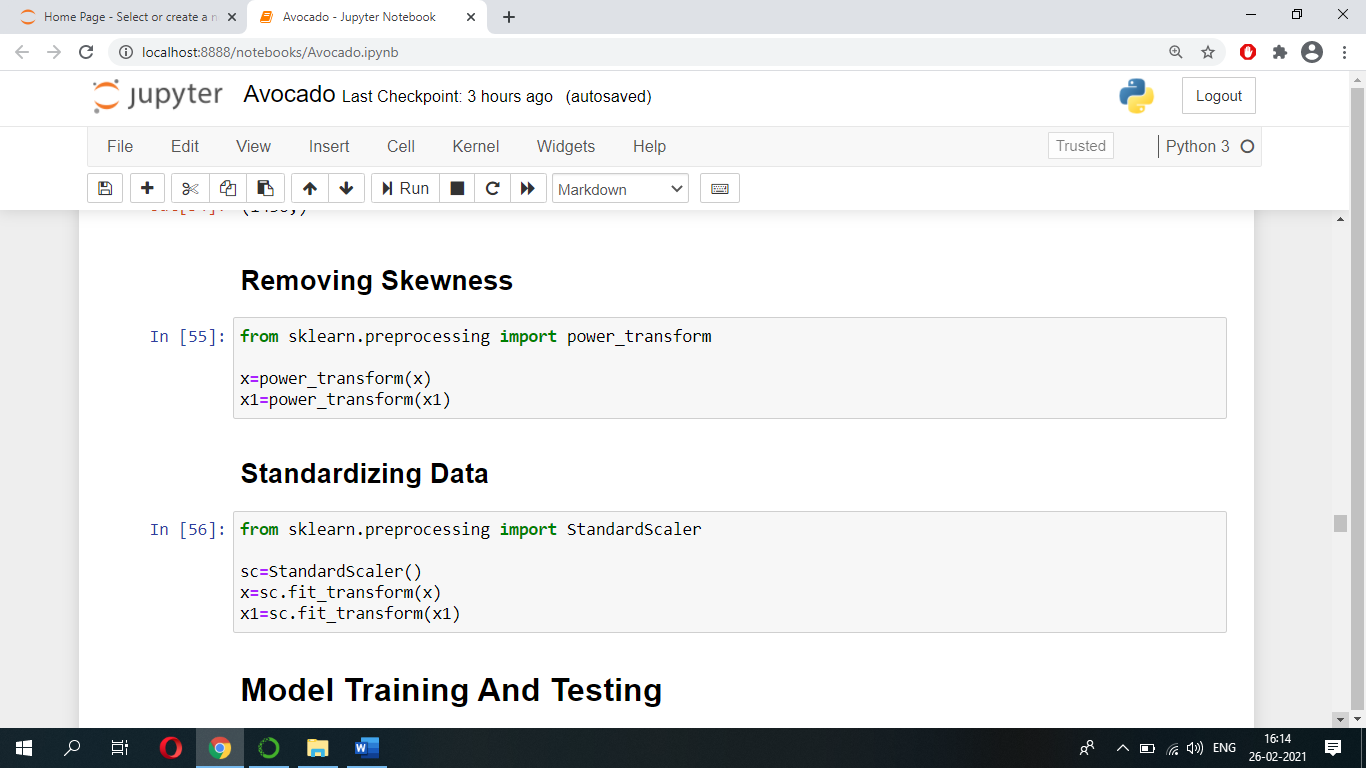
* **Removing skewness**

Now let’s remove skewness by using ‘power transform’ function. We can import ‘power transform’ function from ‘sklearn.preprocessing’.



* **Standardizing Data**

Let’s standardize our data using’ StandardScaler’ function. We can import ‘StandardScaler’ function from ‘sklearn.preprocessing’.

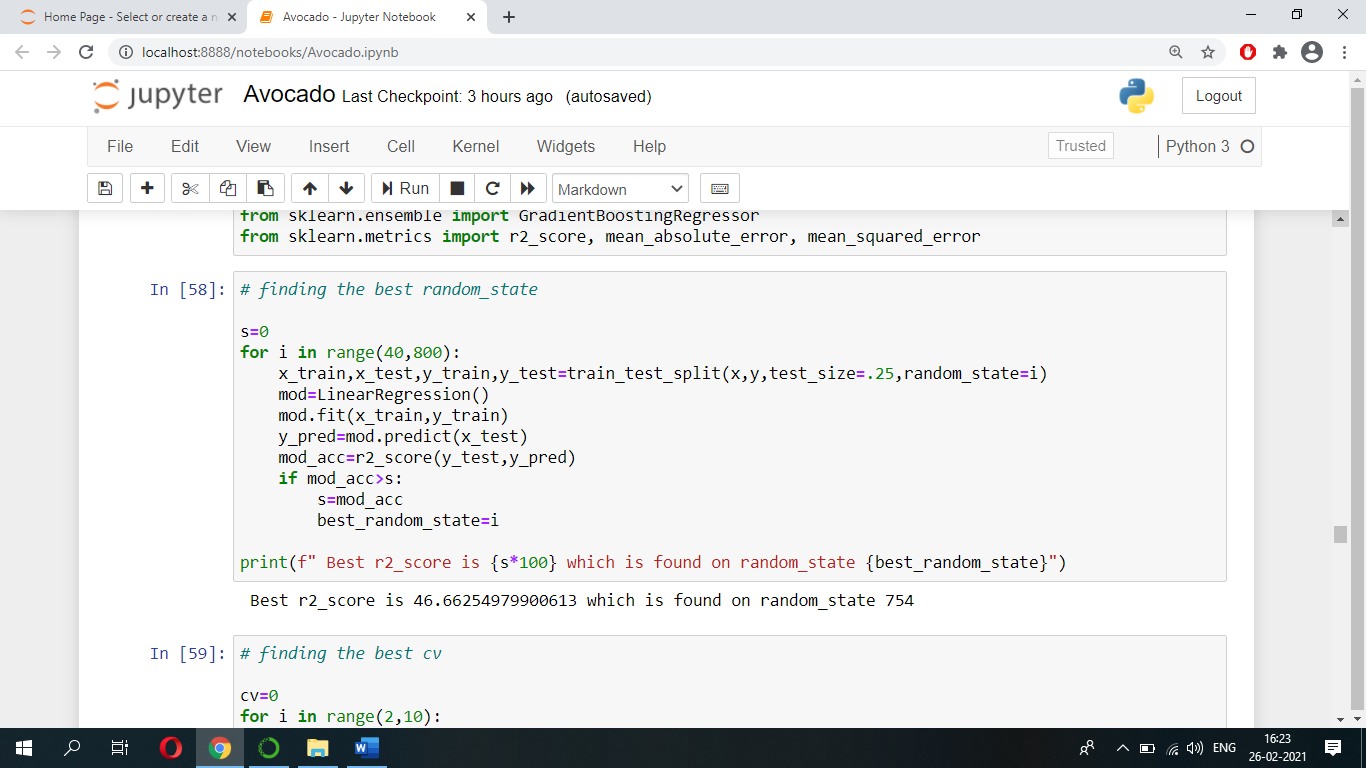


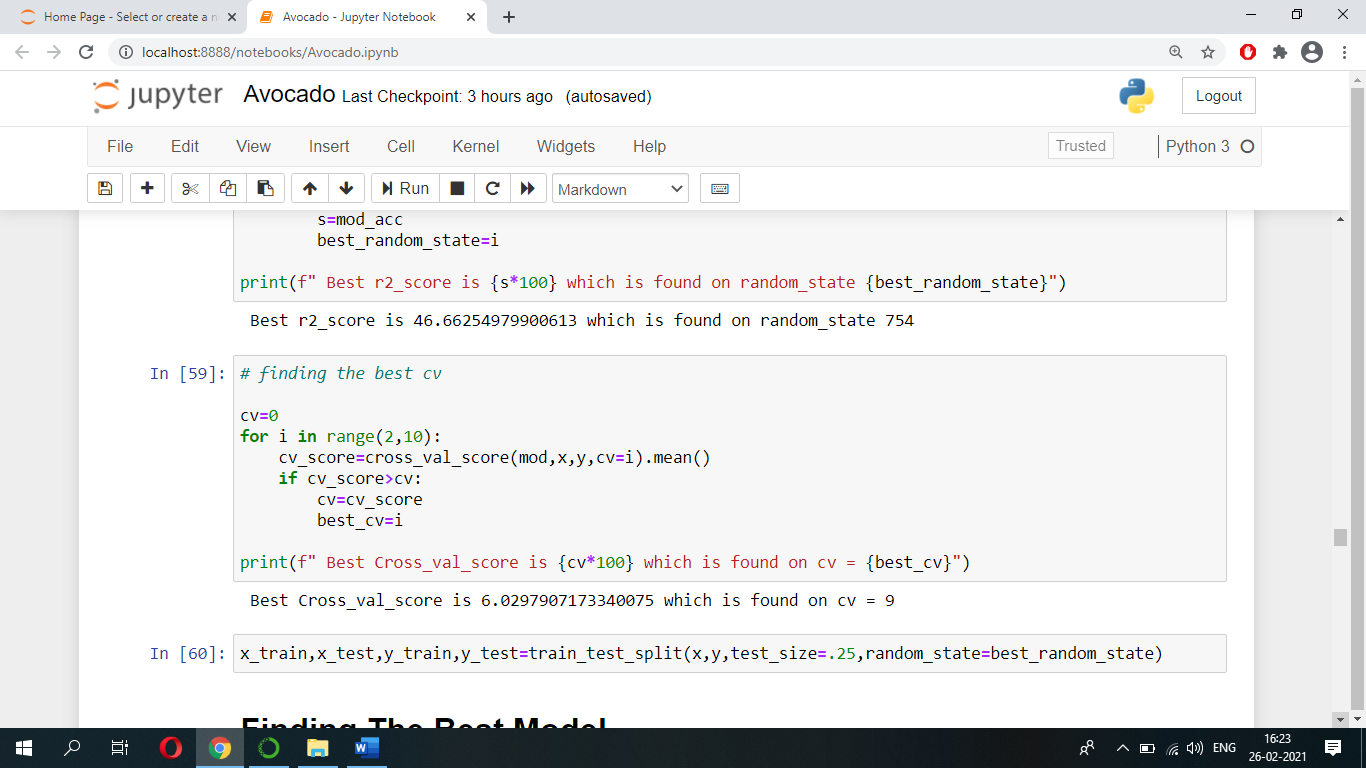
* **Model Training And Testing**

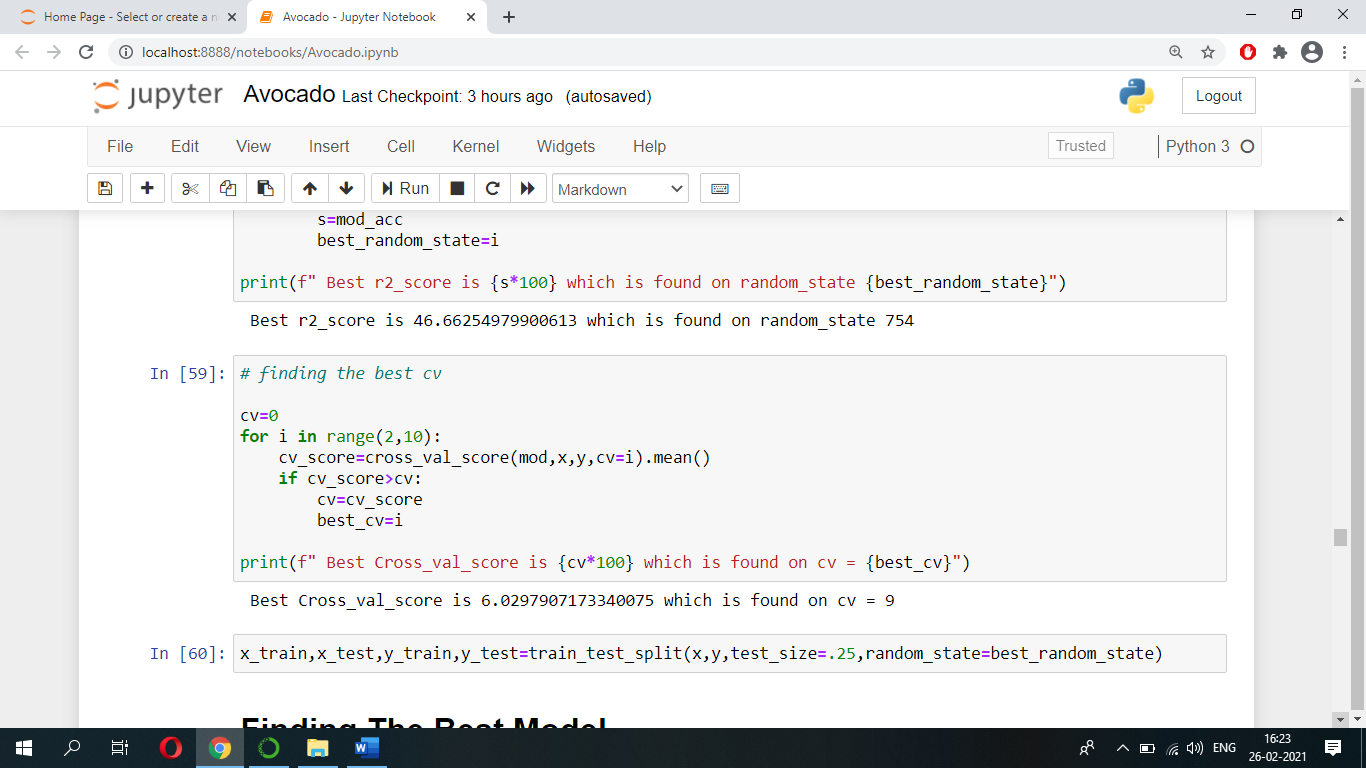
1. **Let’s Predict AveragePrice of an avocado by using Regression Models.**

When using machine learning algorithms/models we should always split our data into a training set and testing set.

Let’s split our data into train set and test set using ‘train\_test\_split’. Let the test size be 25% and remaining 75% be train set.

Below code has been used to find the best random state.

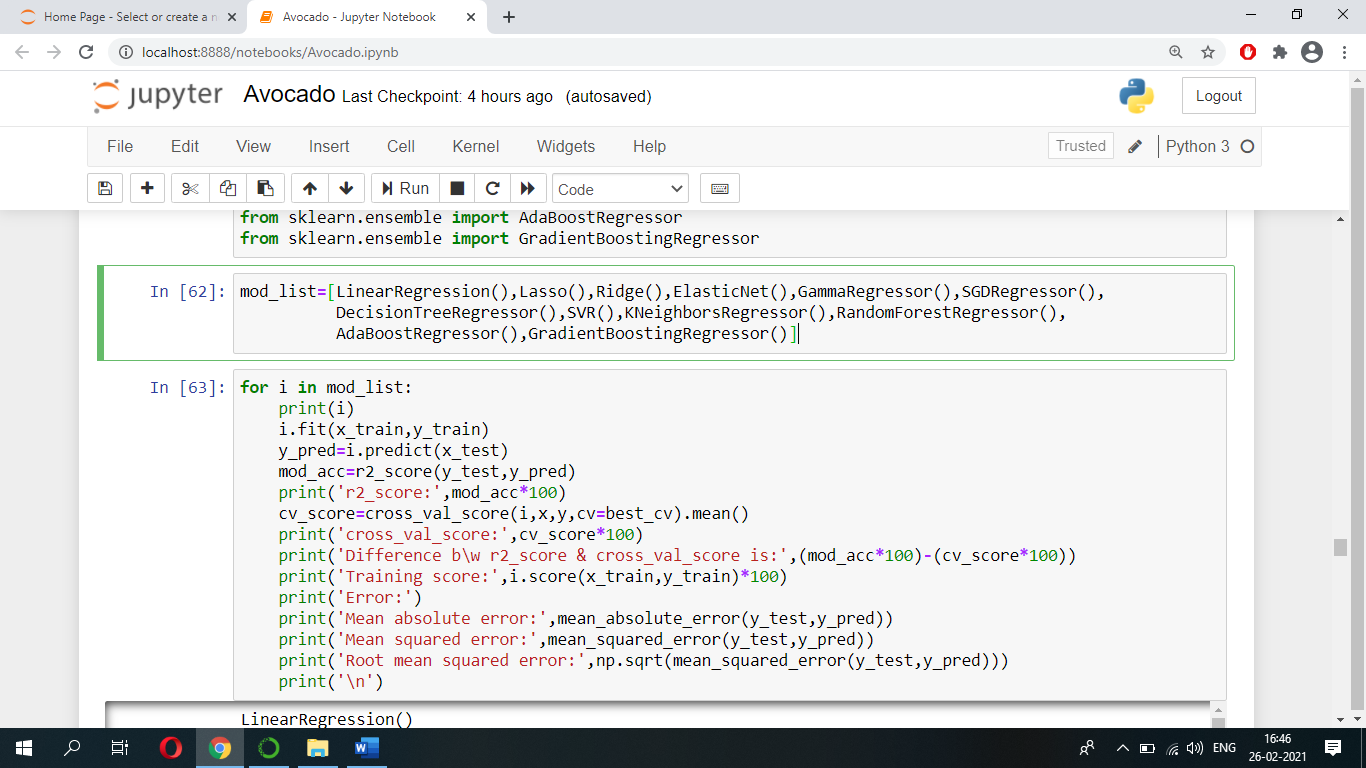
Below code has been used to find the best cv.

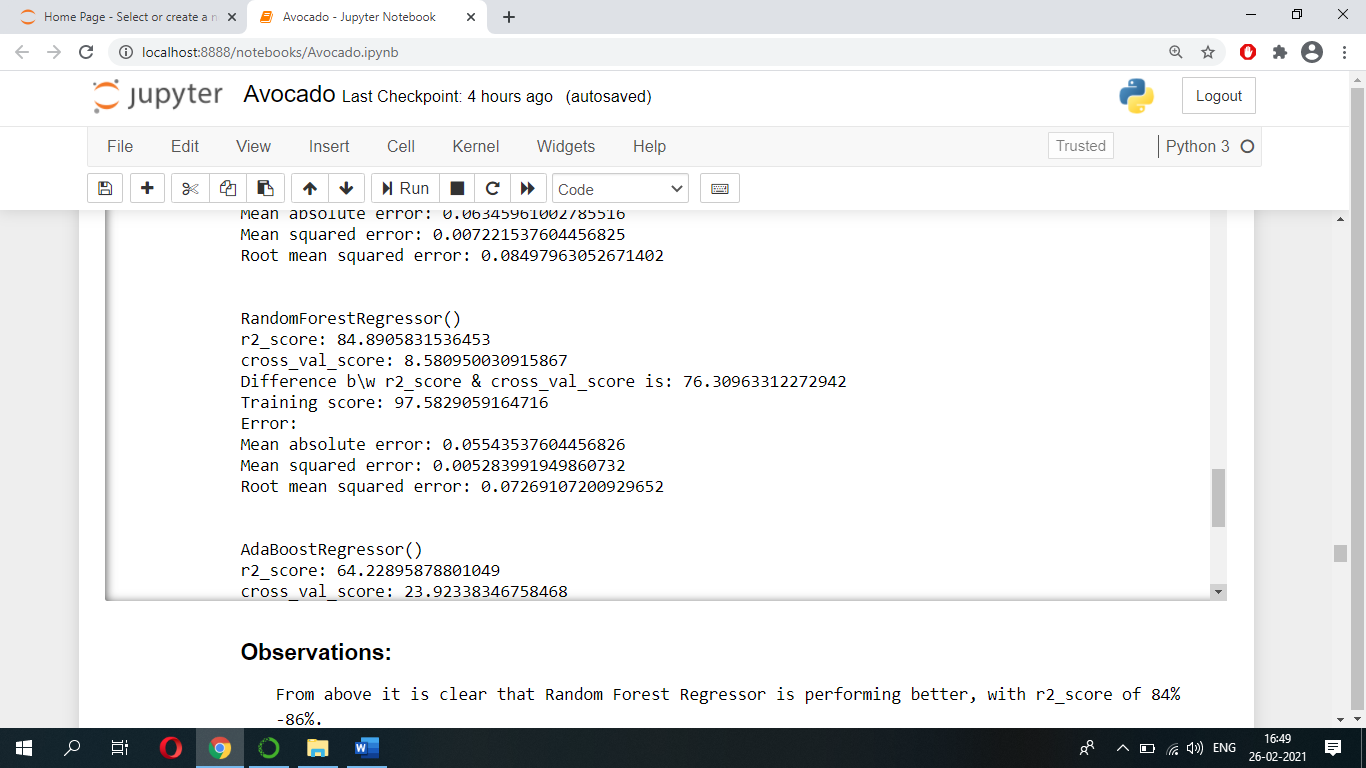
Then the best random state number will be used while splitting our data into train set and test set.

* **Finding The Best Model**

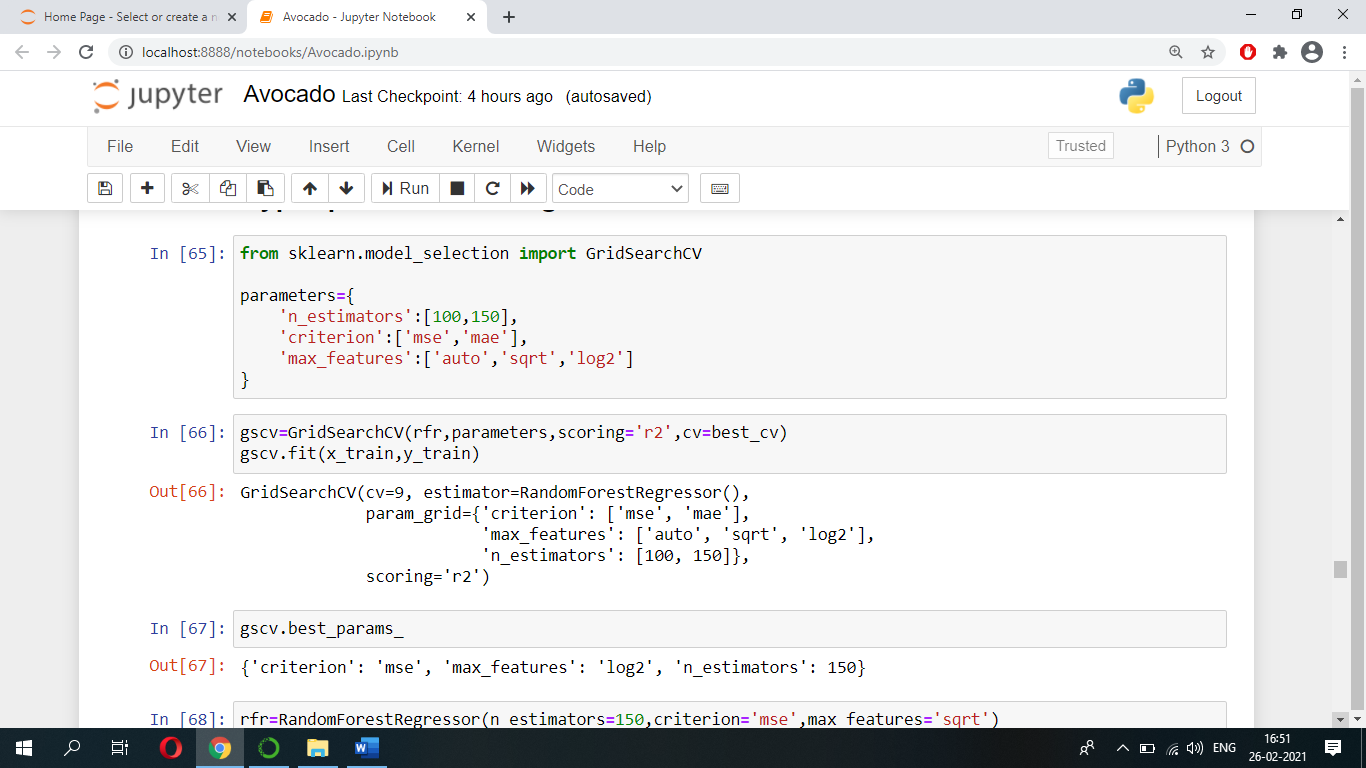
We can now train our model. We use the fit function to train the model.

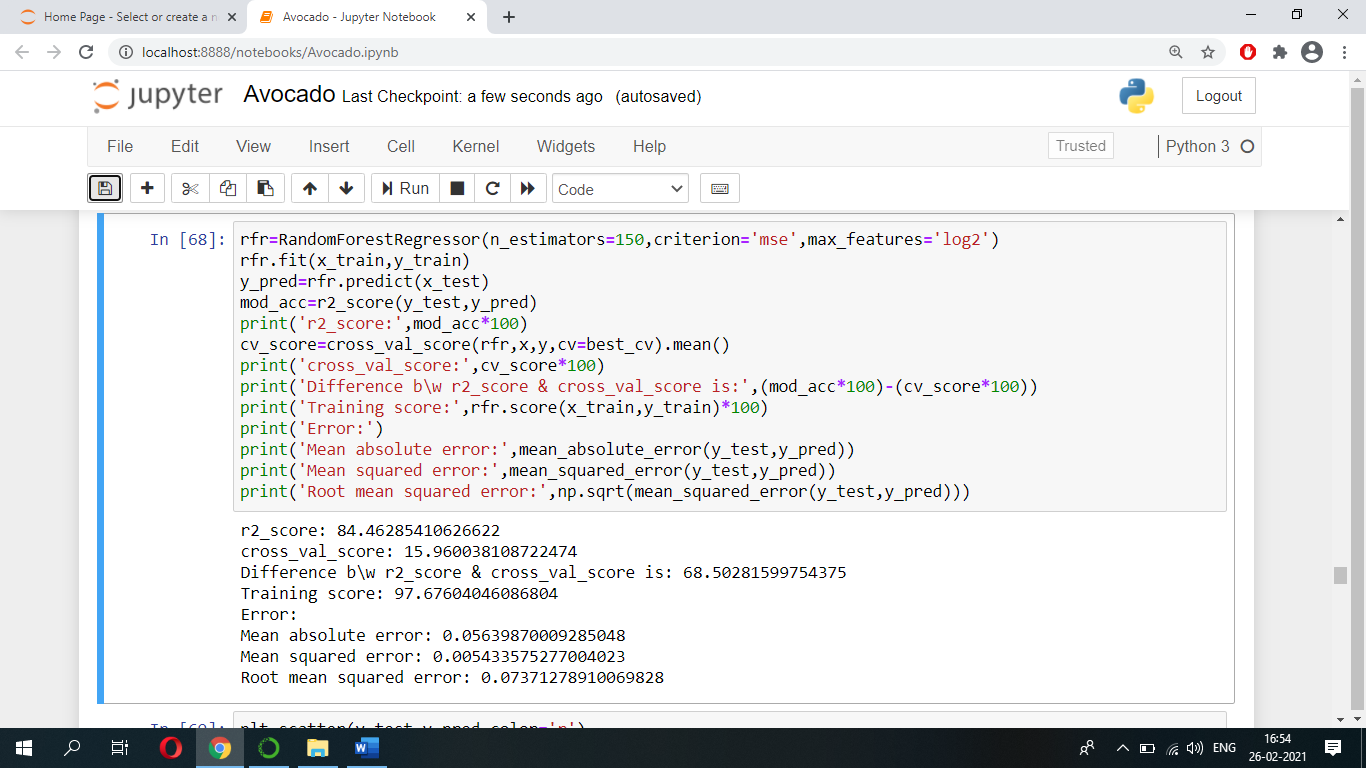
Let’s run the code shown below and see which model gives us high ‘r2\_score’ and less Root mean squared error because model with high r2\_score and less Root mean squared error is a better performing model.



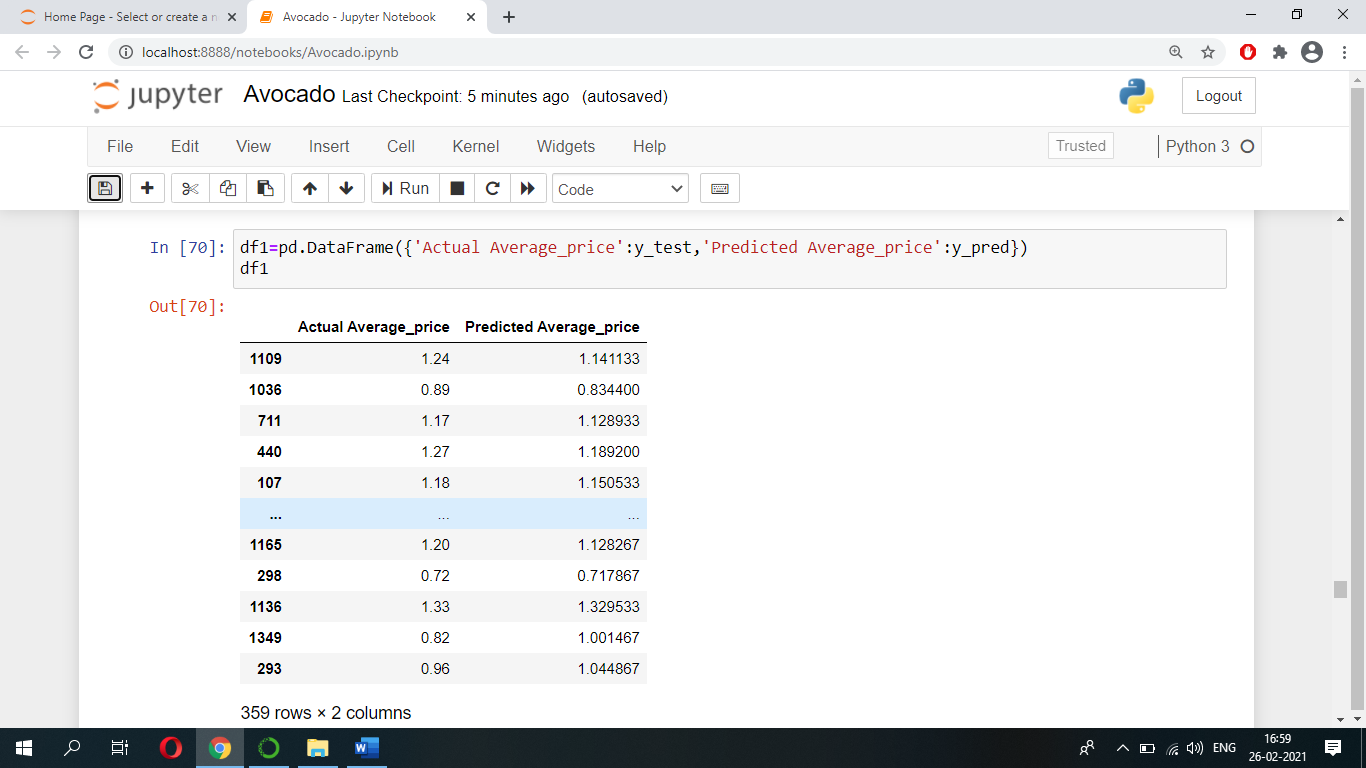
After running the code shown above, we came to know that ‘Random Forest Regressor’ model is performing better with r2\_score of 84 to 86.

Let's perform hyper-parameter tuning on Random Forest Regressor model using GridSearchCV and find out which parameters of Random Forest Regressor model can be used so that model’s performance can be enhanced.

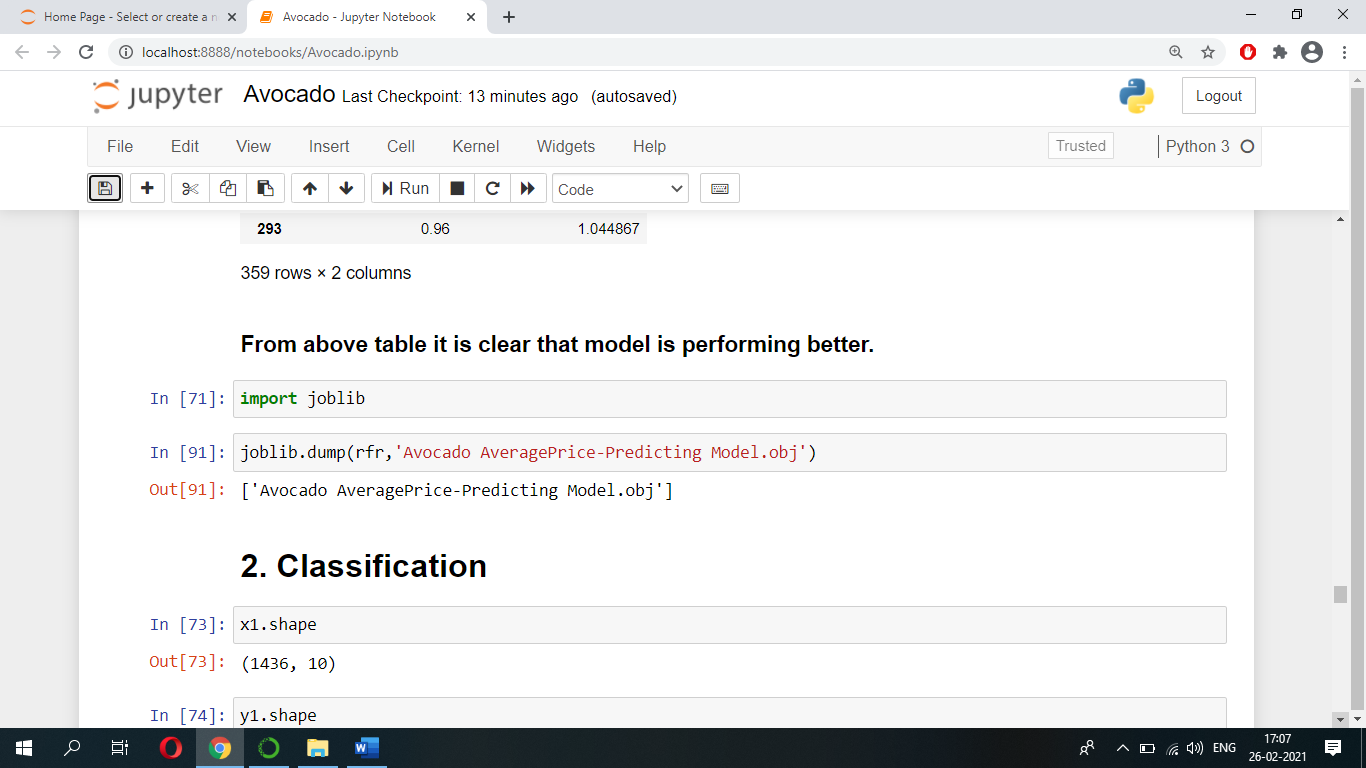
After running below code, we got these parameters which can be used on model to get better performance.

**Let’s use these parameters and see what r2\_score we are going to get.

Finally, we got r2\_score of 84.46.

We can see how our model is performing i.e., predicting AveragePrice of an avocado from below table.

Lastly, let’s save this model for production and for future prediction using dump function of joblib.

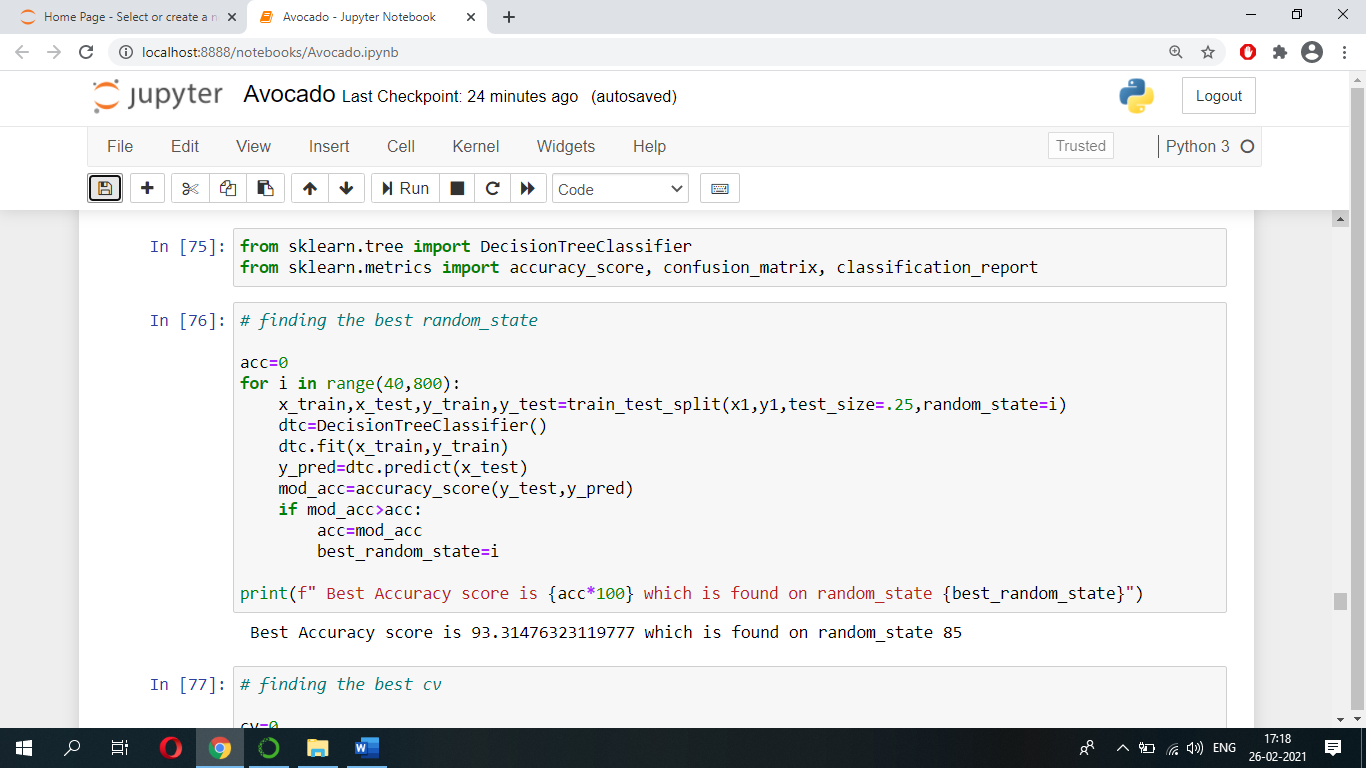


1. **Let’s Predict regions/cities by using Classification Models.**

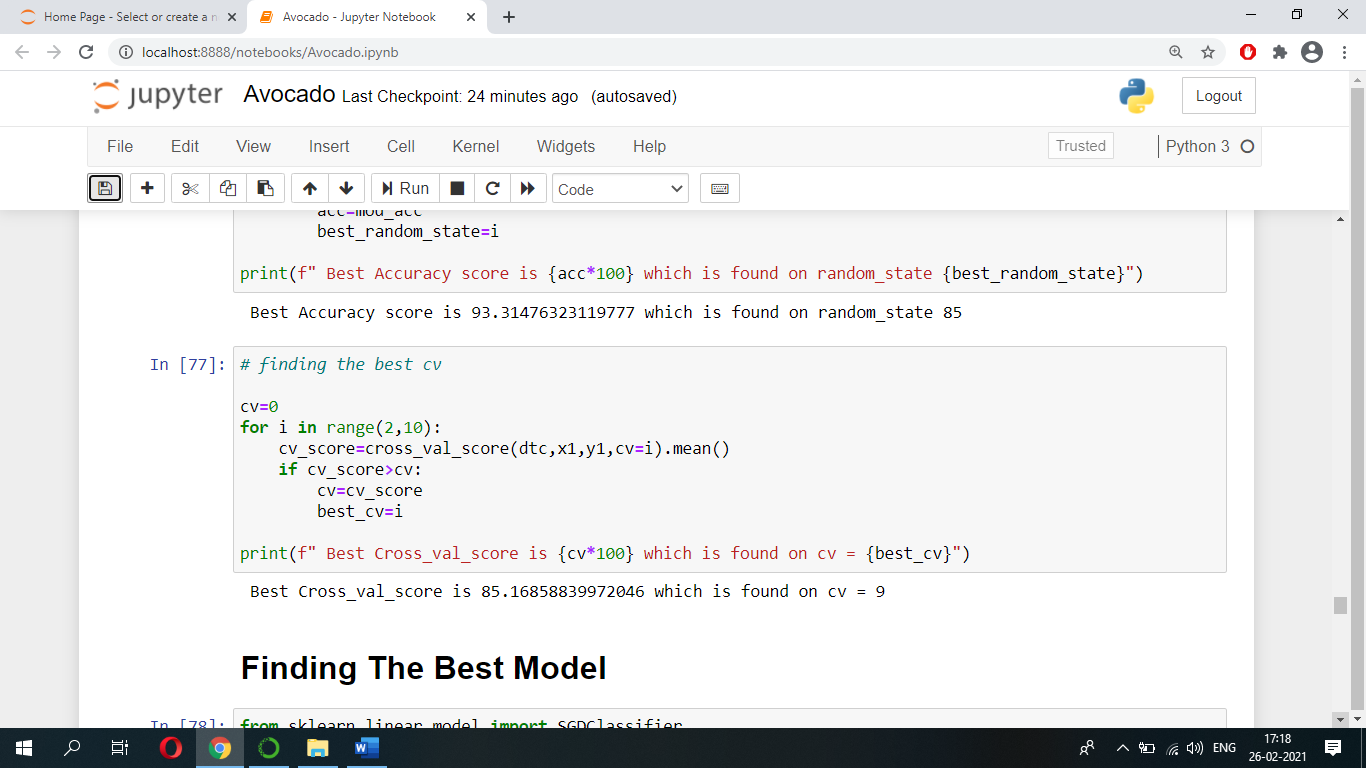
When using machine learning algorithms/models we should always split our data into a training set and testing set.

Let’s split our data into train set and test set using ‘train\_test\_split’. Let the test size be 25% and remaining 75% be train set.

Below code has been used to find the best random state.



Below code has been used to find the best cv.



Then the best random state number will be used while splitting our data into train set and test set.

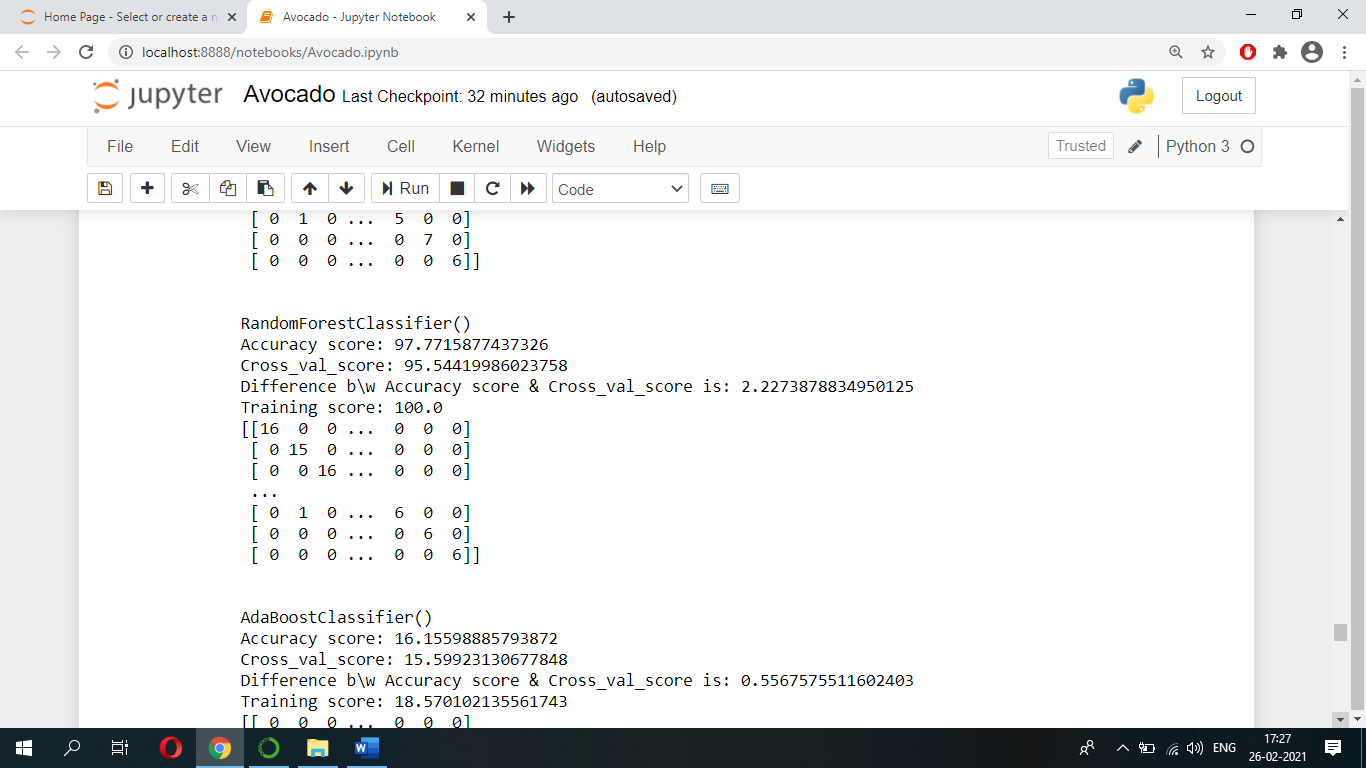


* **Finding The Best Model**

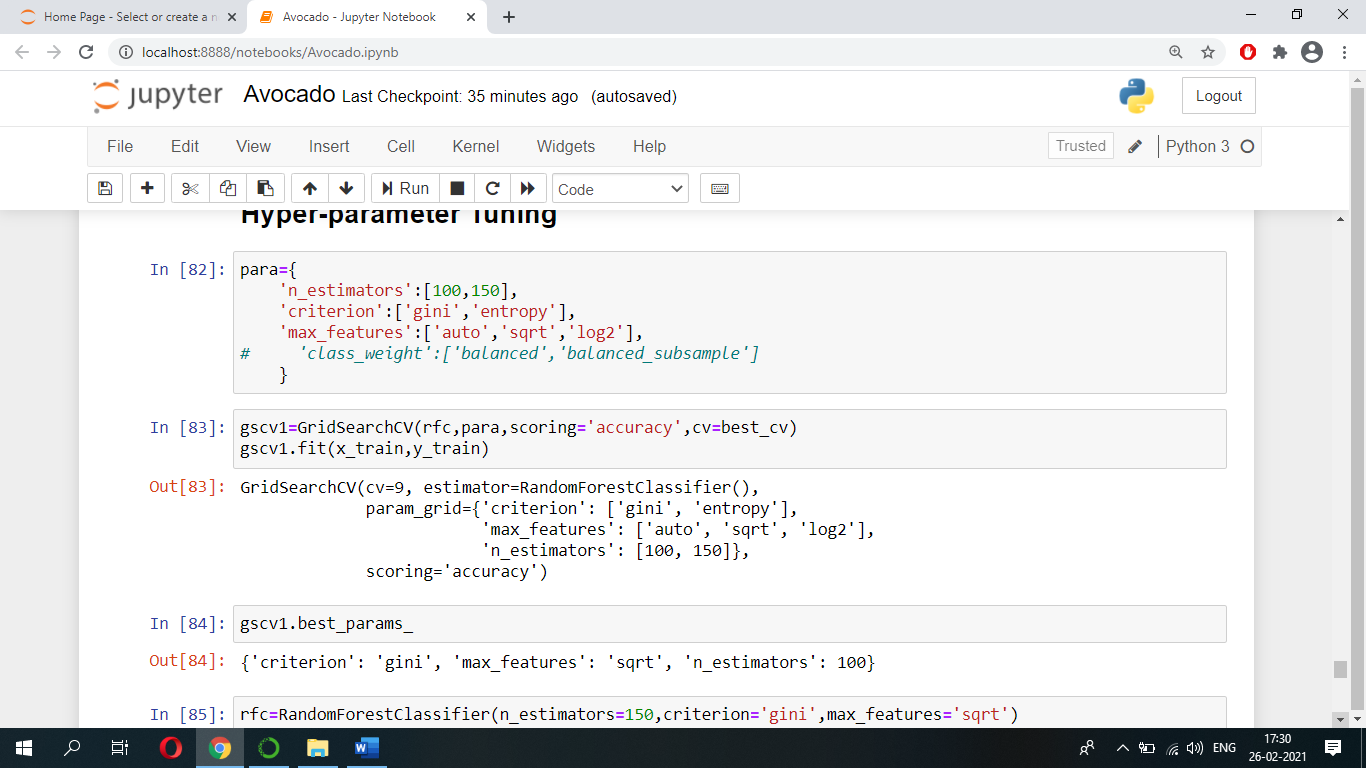
We can now train our model. We use the fit function to train the model.

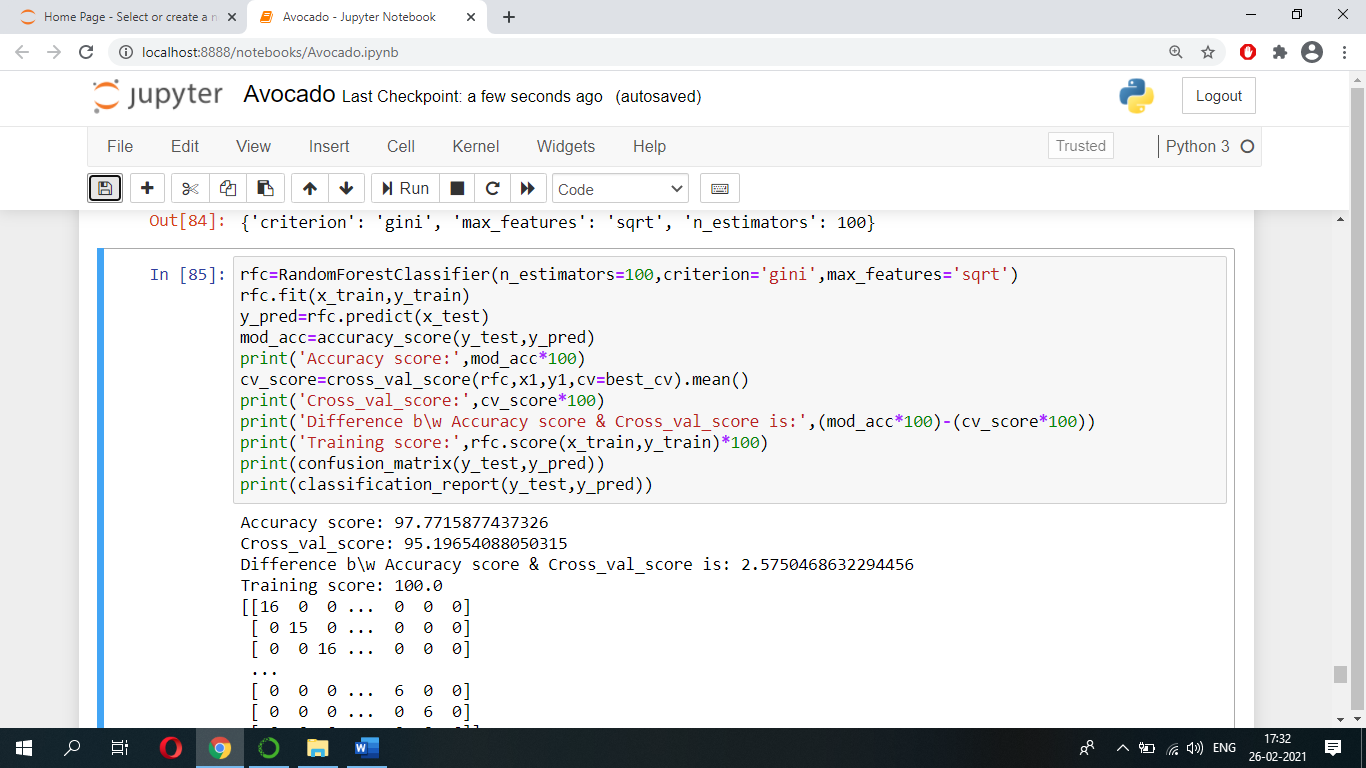
Let’s run the code shown below and see which model gives us high ‘accuracy score’ because higher the accuracy score better is the model performance.



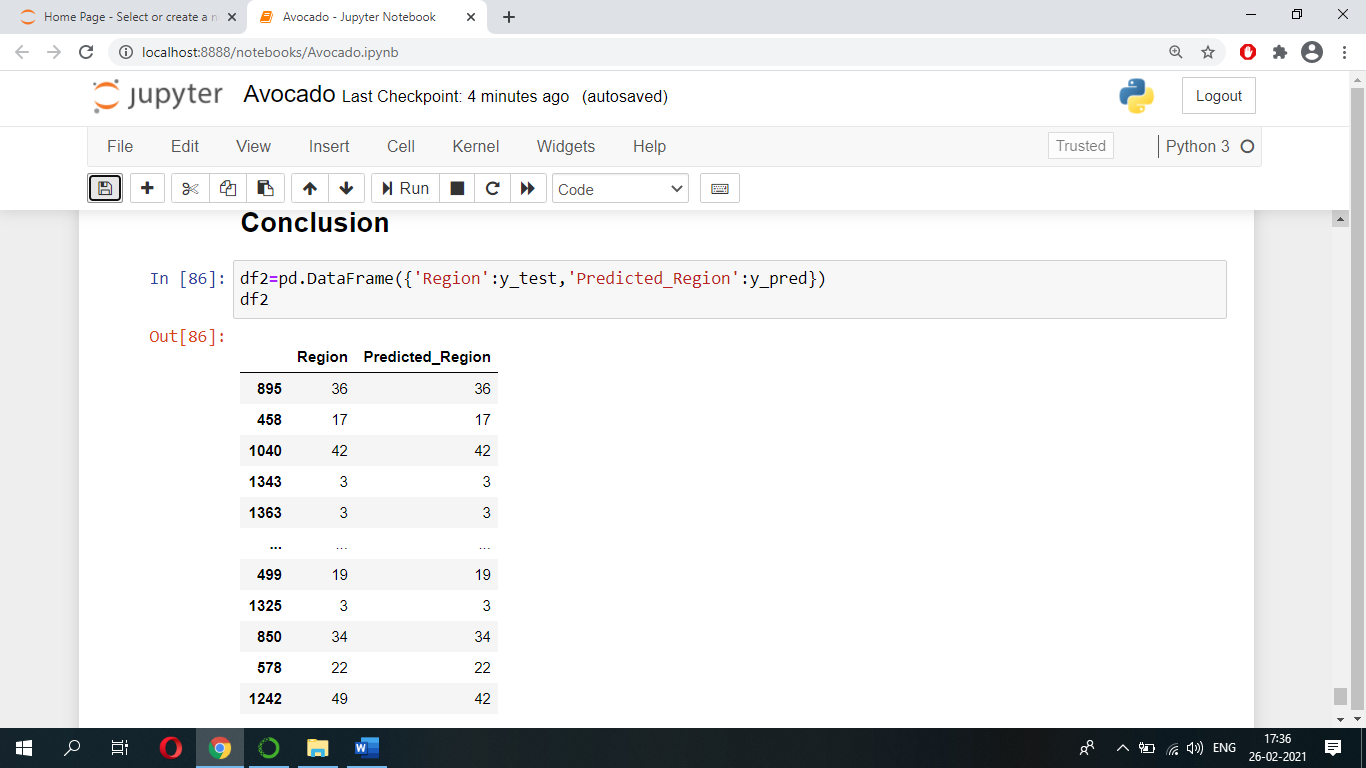
After running the code shown above, we came to know that ‘Random Forest Classifier’ model is performing better with accuracy score of 96 to 98.

Let's perform hyper-parameter tuning on Random Forest Classifier model using GridSearchCV and find out which parameters of Random Forest Classifier model can be used so that model’s performance can be enhanced.

After running below code, we got these parameters which can be used on model to get better performance.

Let’s use these parameters and see what accuracy score we are going to get.

Finally, we got accuracy score of 97.77.

We can see how our model is performing i.e., predicting regions/cities from below table.

Lastly, let’s save this model for production and for future prediction using dump function of joblib.

