**Data Science Home Price Estimator:**

* Created a tool that estimates home price in India to help people negotiate the price when they get a house.

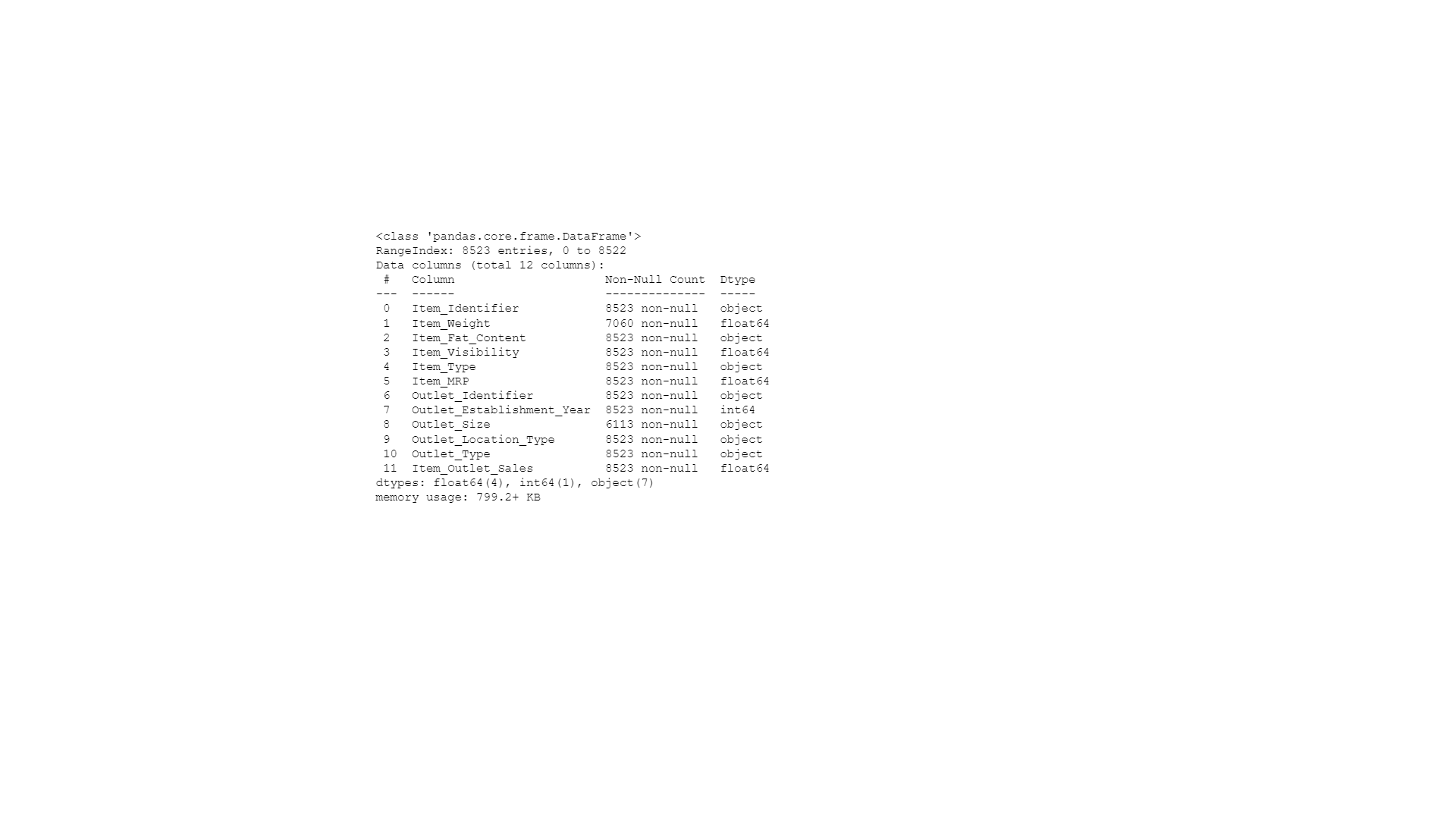
**Packages:** pandas, numpy, sklearn, matplotlib, seaborn,

**The solving mechanism**

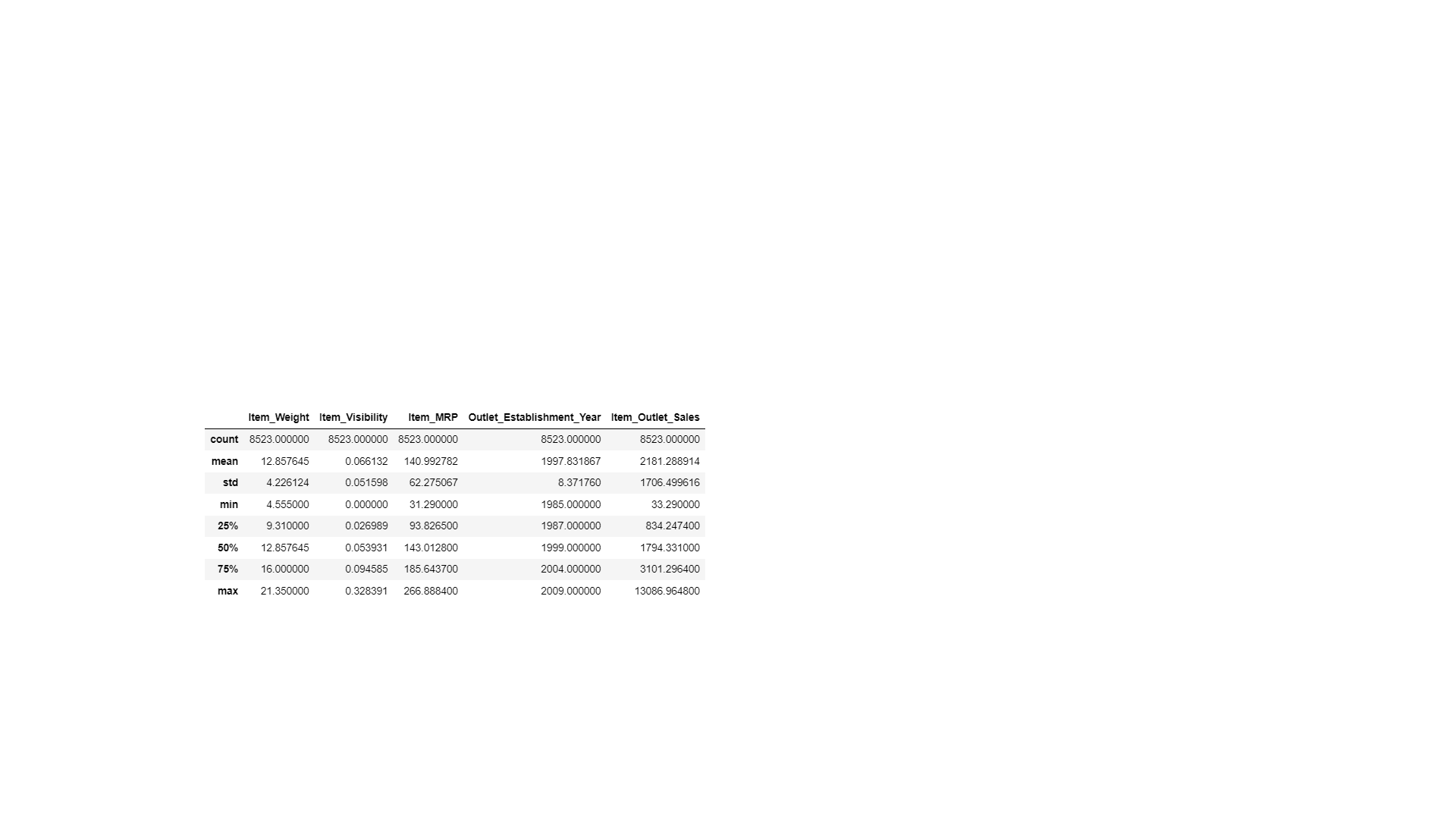
* build machine learning model using python

**Describe the dataset**

* Data source:
  + <https://www.kaggle.com/datasets/amitabhajoy/bengaluru-house-price-data>
* Data description
* I use pandas library to description dataset
  + big\_mart\_data.info()
  + Output:



* + form output I know number of rows (13320) and num of columns (9)
  + name of columns and data type for each column
  + number of null values in columns (Ex: bath has 53 sell null)
  + big\_mart\_data.describe()
  + output:



* + I conclude from this table count , mean , min , median , max , standard deviation
  + From this information I know count of value in each column (EX: balcony has 12711

Value which means 609)

* + Std mean standard deviation it help us to know the spread of values
  + Max , Min , mean , Median of each column

**Data Cleaning and EDA**

After collection the data, I needed to clean it up so that it was usable for our model. I made the following changes and created the following variables:

* **Fill the column (Item\_Weight) with mean value**
* **Fill the column (Outlet\_Size) with mode value**
* **Data Pre-Processing in Item\_Fat\_Content column**

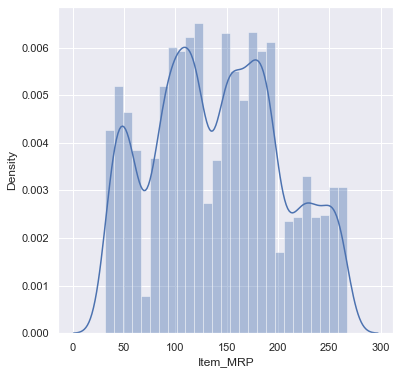
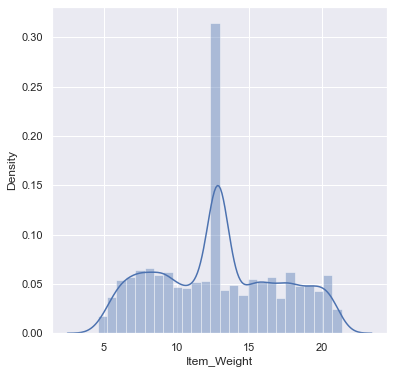
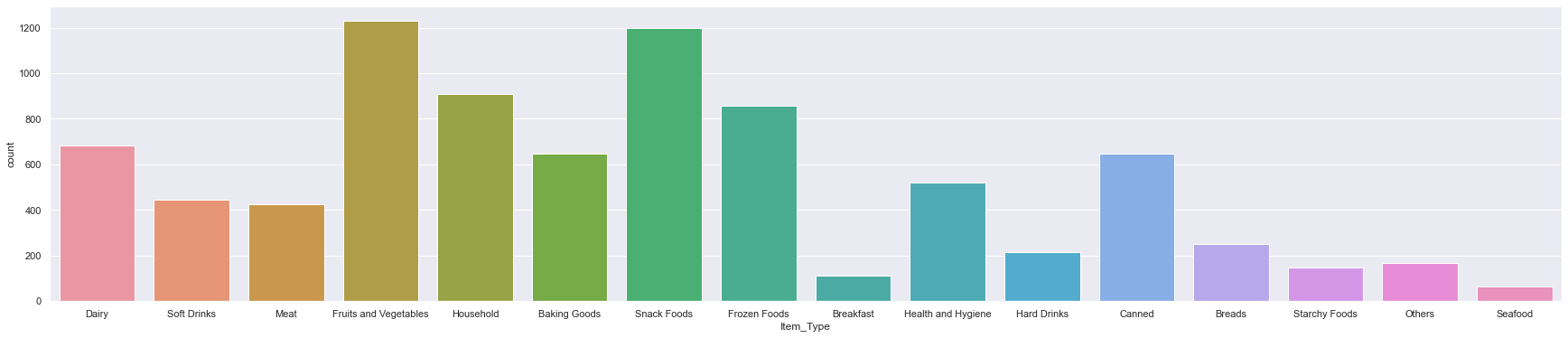
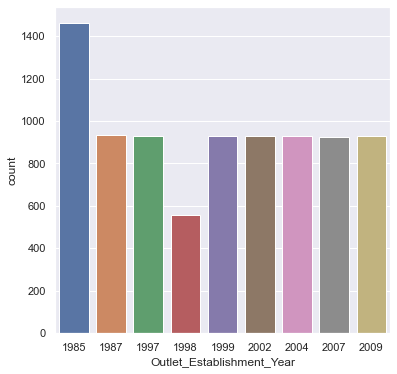
**Label Encoding in all columns categories [**

1. Item\_Identifier
2. Item\_Fat\_Content
3. Item\_Type
4. Outlet\_Identifier
5. Outlet\_Size ,
6. Outlet\_Location\_Type ,
7. Outlet\_Type

**]**

**descriptive statistics and data distribution charts**

I looked at the distributions of the data and the value counts for the various categorical variables. Below are a few highlights from the pivot tables.



**التعليق**

* في المحور X يوضح مقدار تنوع الاسعار في data set و في y مقدار تكرار الاسعار
* في المحور X يوضح مقدار تنوع حجم المنتجات في data set و في y مقدار تكرار حجم المنتجات

**Model Building**

spilt data to x and y

I also split the data into train and tests sets with a test size of 20%.

I tried model:

* **Linear Regression**– Baseline for the model

**Model performance**

The Linear Regression model outperformed in the test and validation.

* **Linear Regression**: = 72.52%