DATA MINING SWE 2007

REVIEW -2

**TEAM MEMBERS:**

DEEPIKA G - 19MIS1159

BHUVANESHWARI K - 19MIS1196

GURU PRASHATH M – 19MIS1122

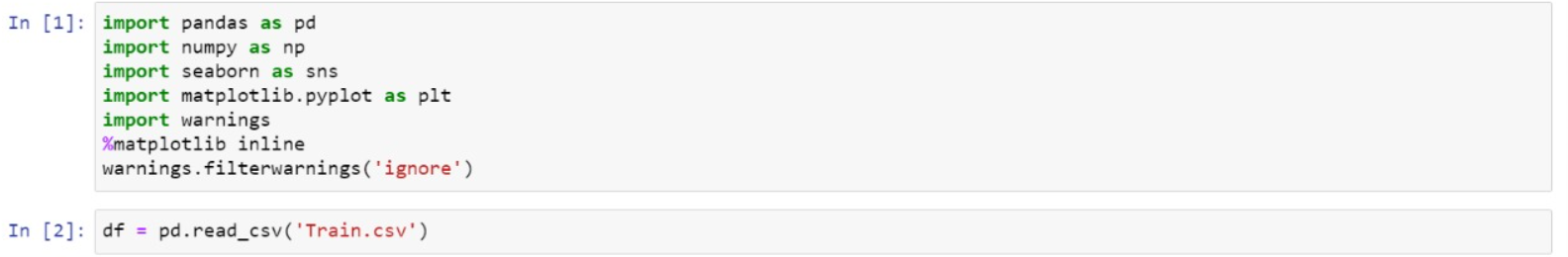
**PROJECT TITLE: BIG MARKET SALES PREDICTION**

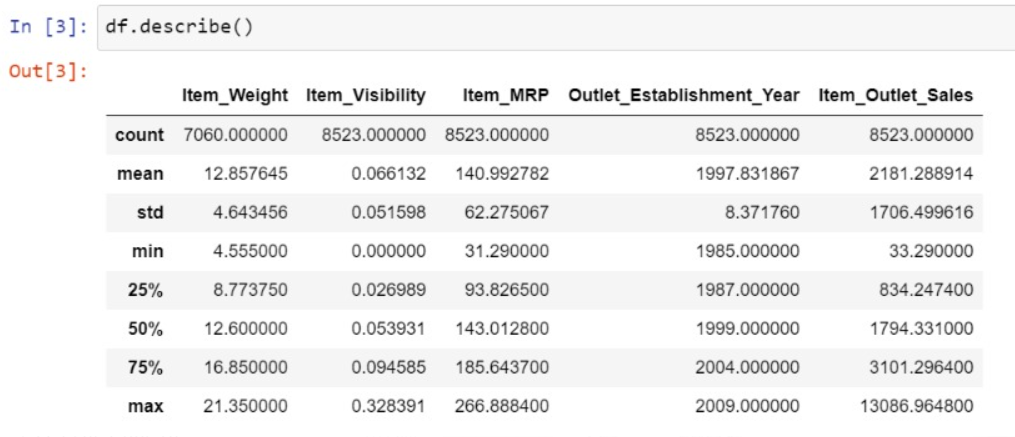
**IMPLEMENTATION:**

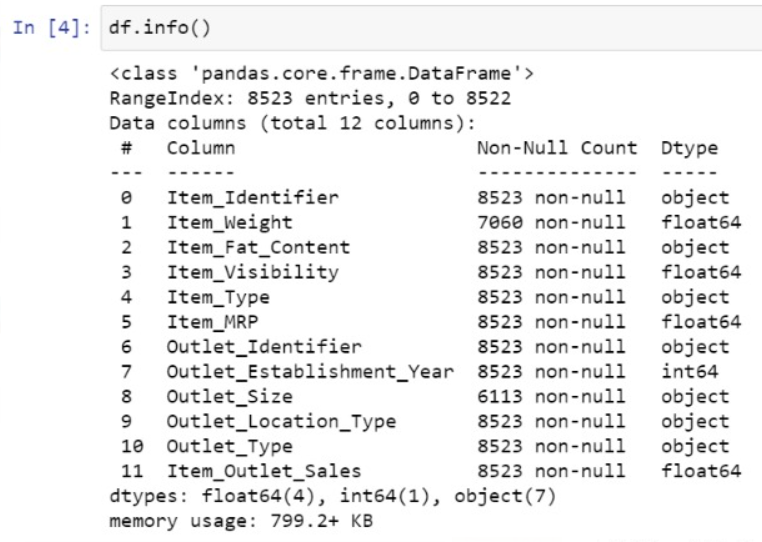
**PREPROCESSING DATA:**

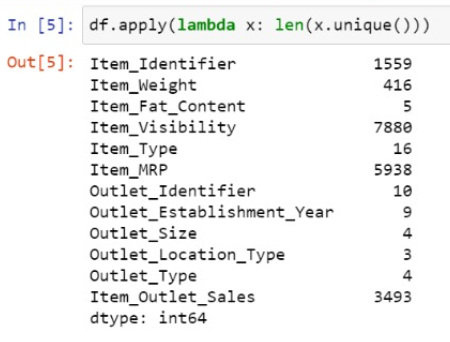
Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data pre-processing task.

Screenshots of code.

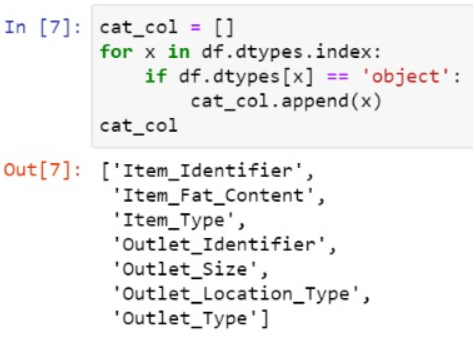
****

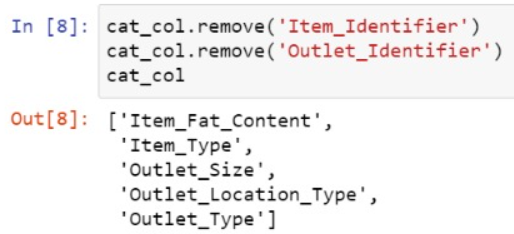
****

****

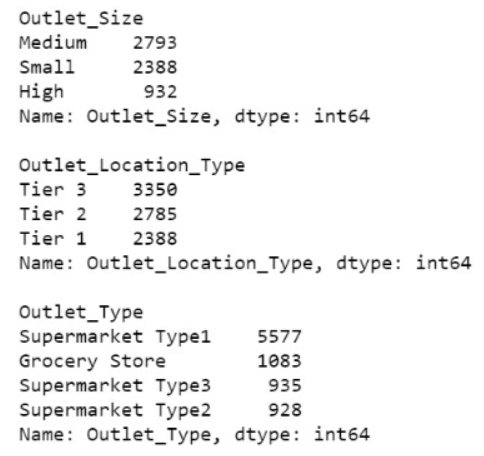
****

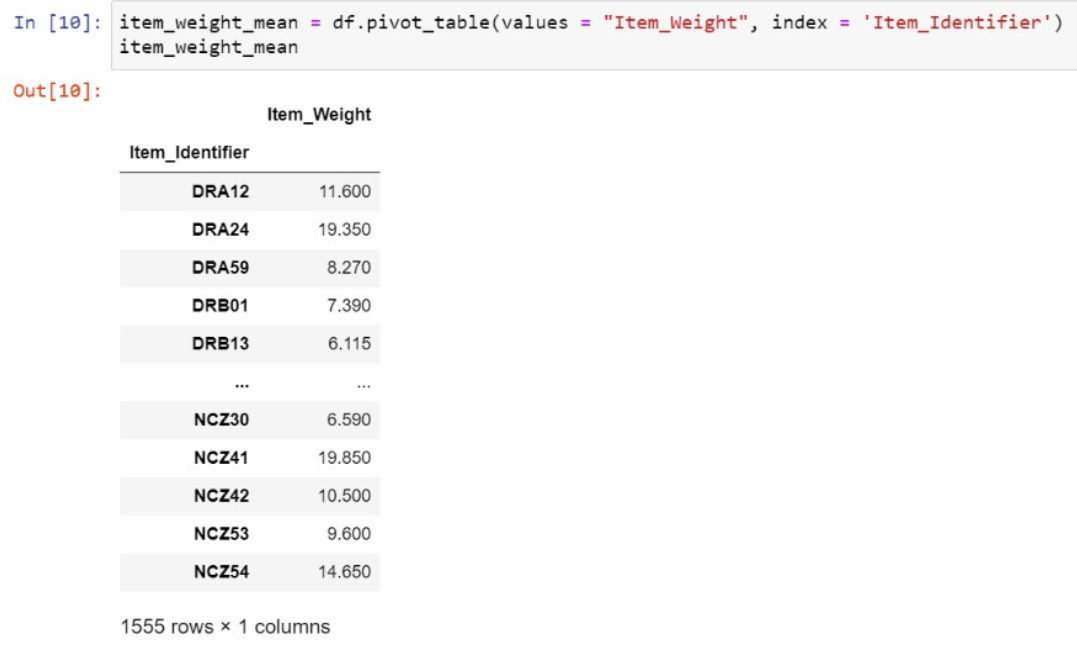
****

****

****

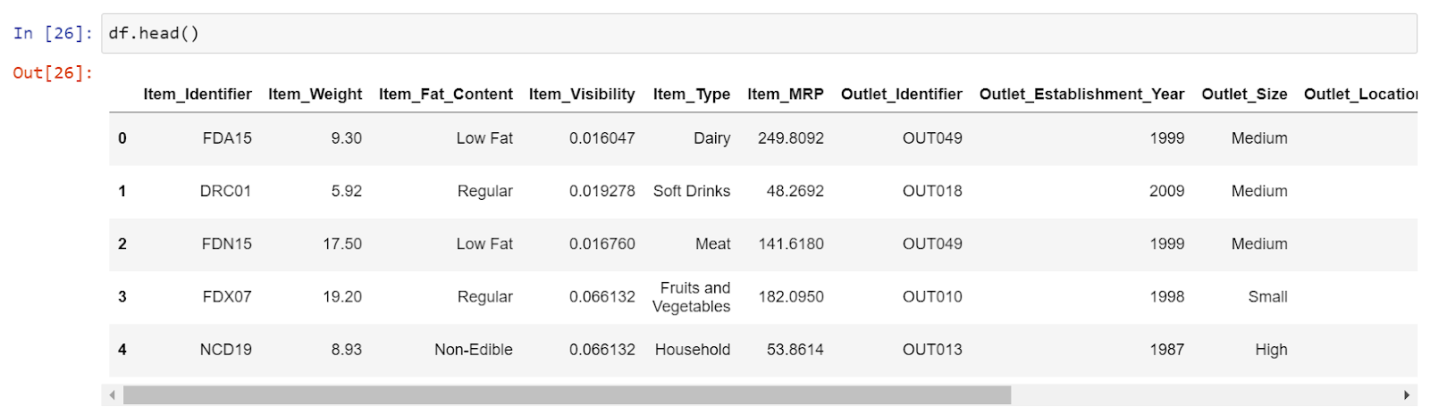
****

****

****

****

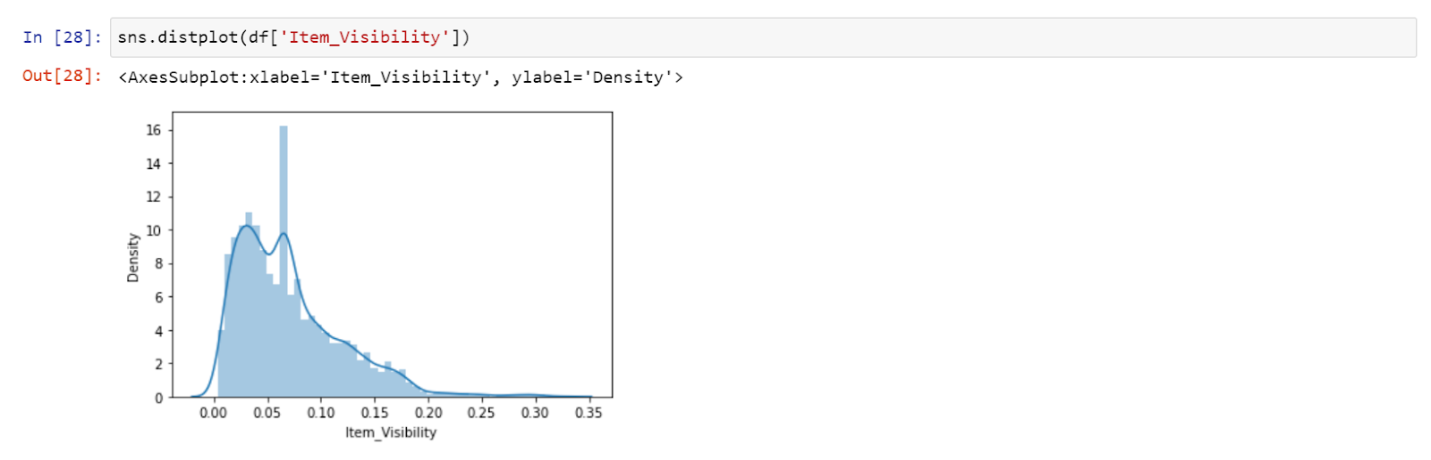




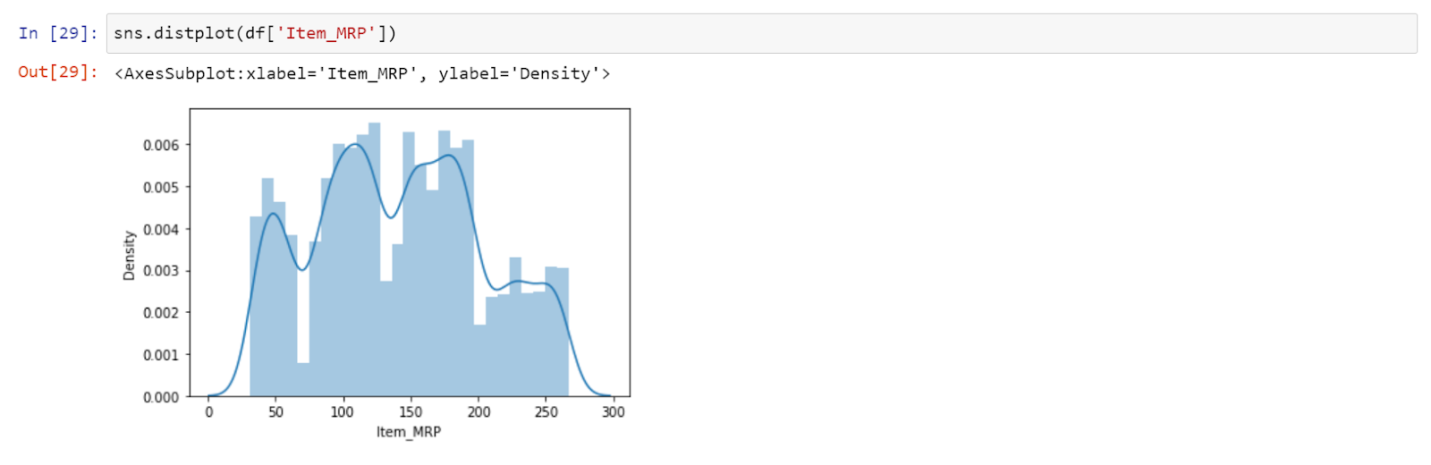
**DATA VISUALIZATION:**

Data visualization is the representation of data or information in a graph, chart, or other visual format. It communicates relationships of the data with images. This is important because it allows trends and patterns to be more easily seen. With the rise of big data upon us, we need to be able to interpret increasingly larger batches of data. Machine learning makes it easier to conduct analyses such as predictive analysis, which can then serve as helpful visualizations to present. But data visualization is not only important for data scientists and data analysts, it is necessary to understand data visualization in any career.

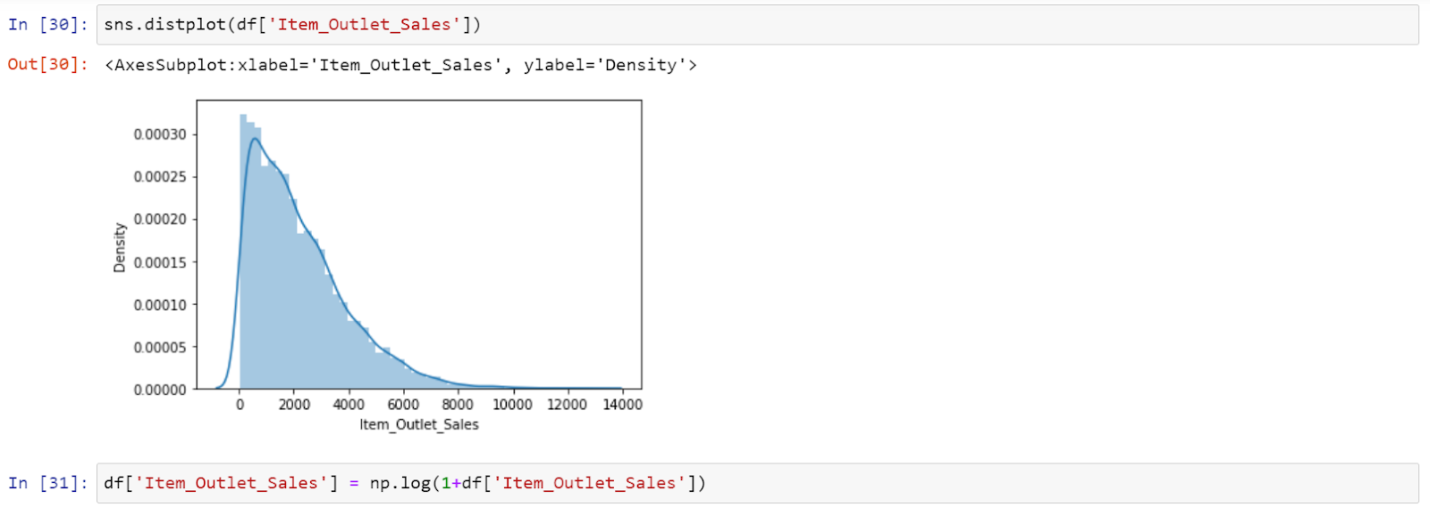
Data visibility graph

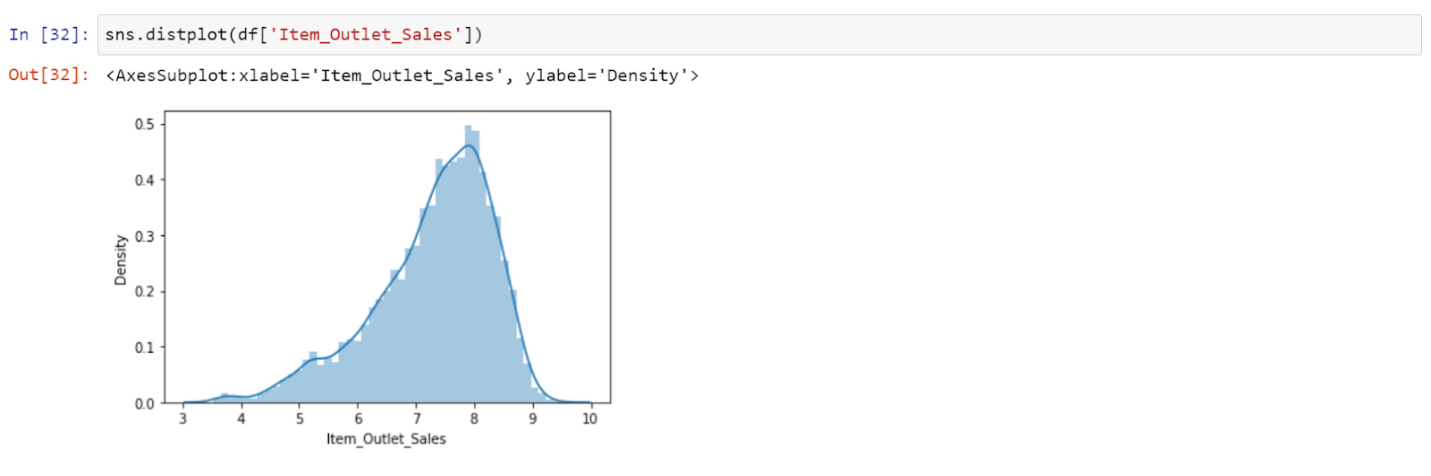


Item MRP graph

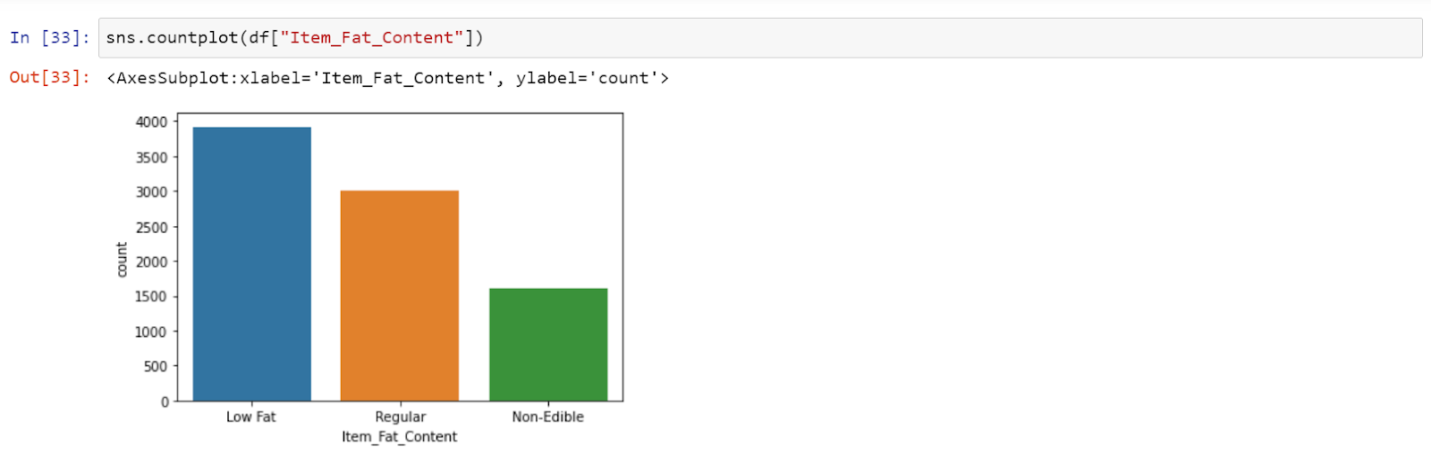


Item outlet sale graph

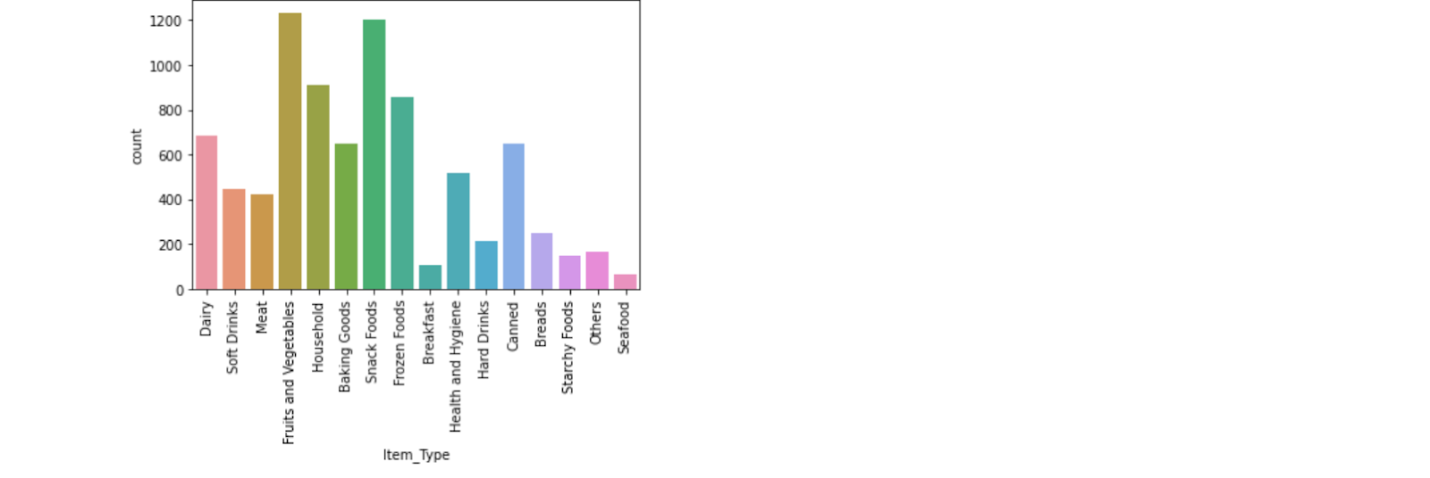
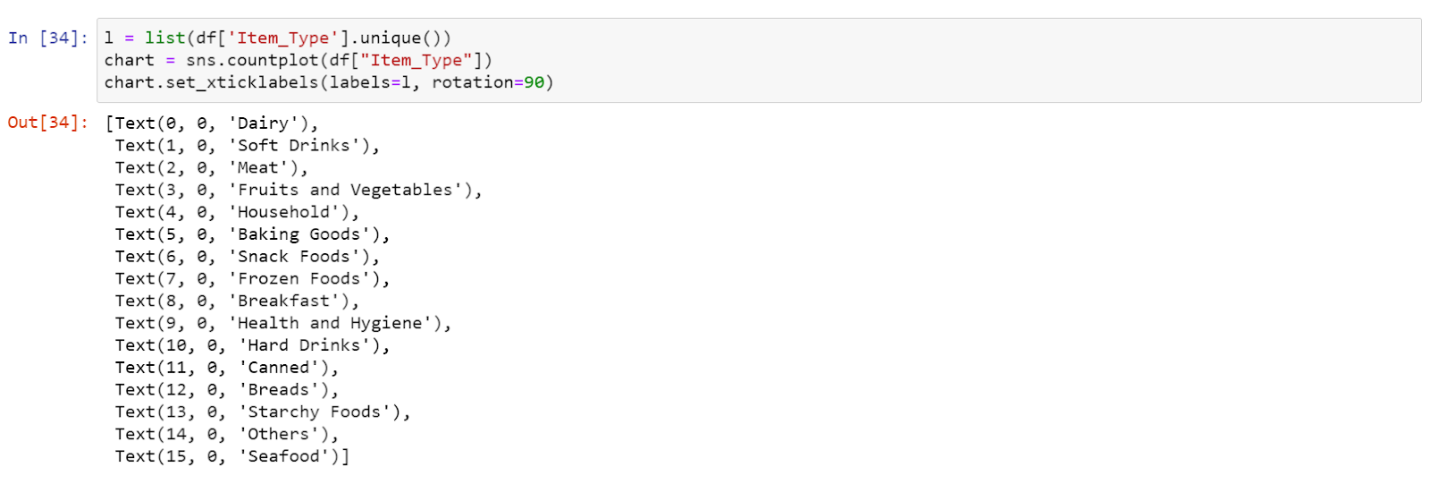


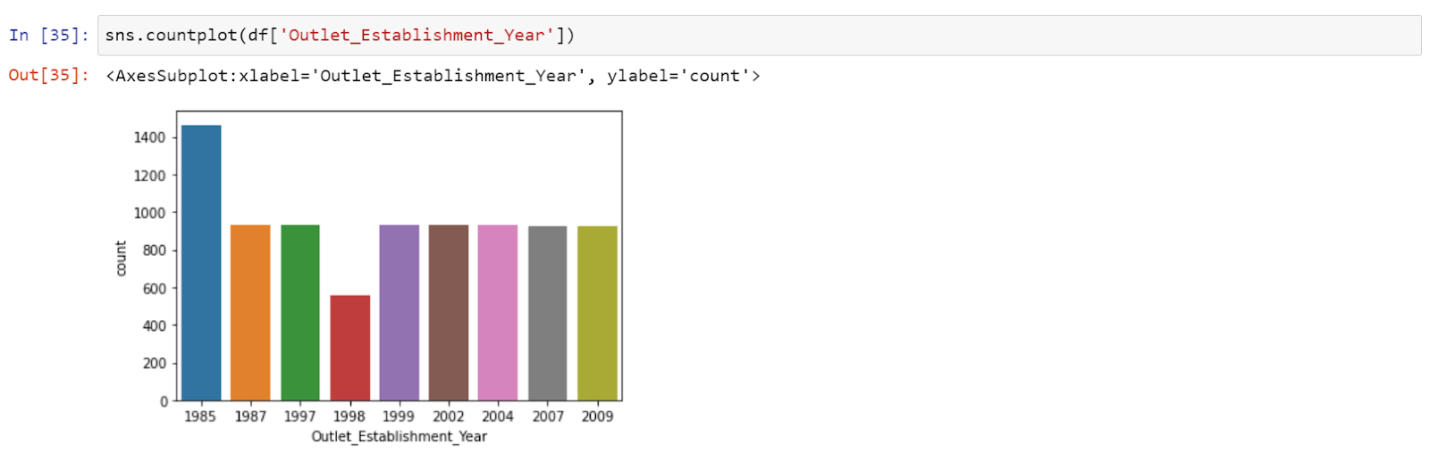
Item outlet sales graph after applying log

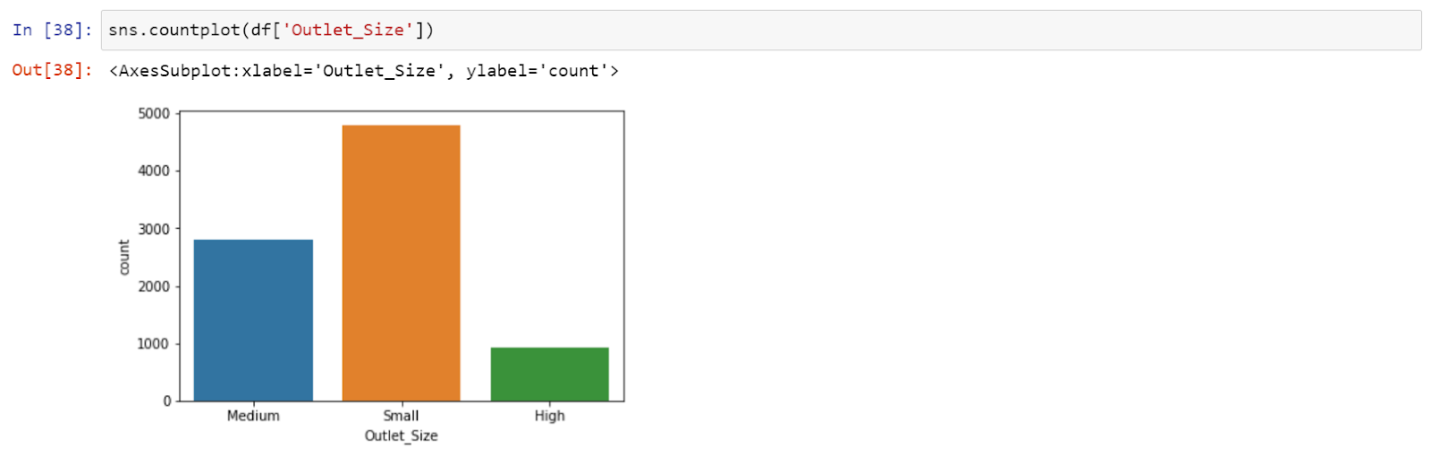
Item fat content bar graph

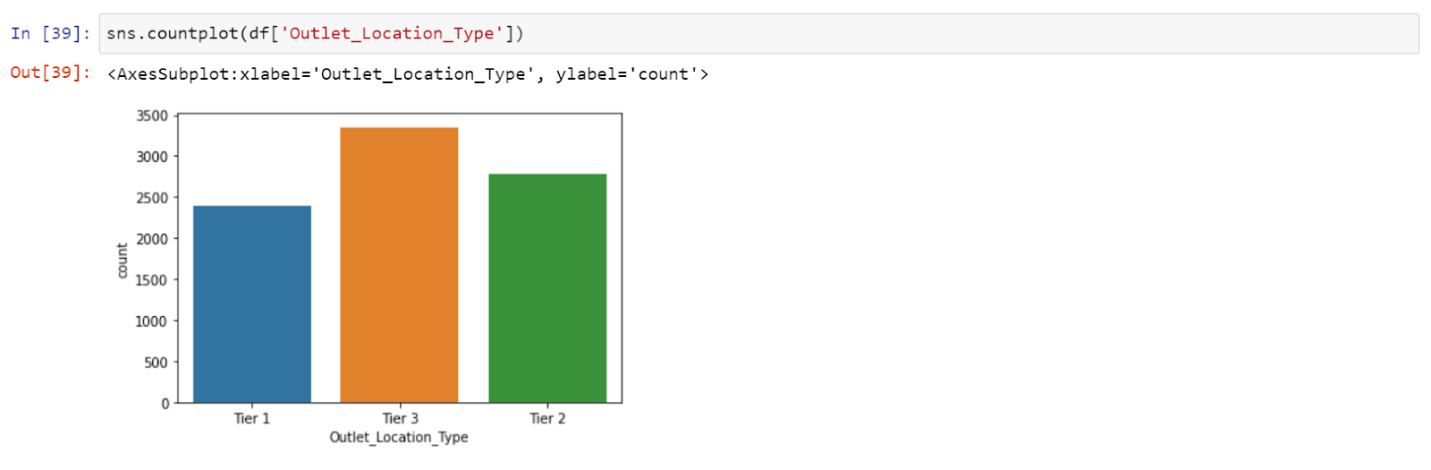


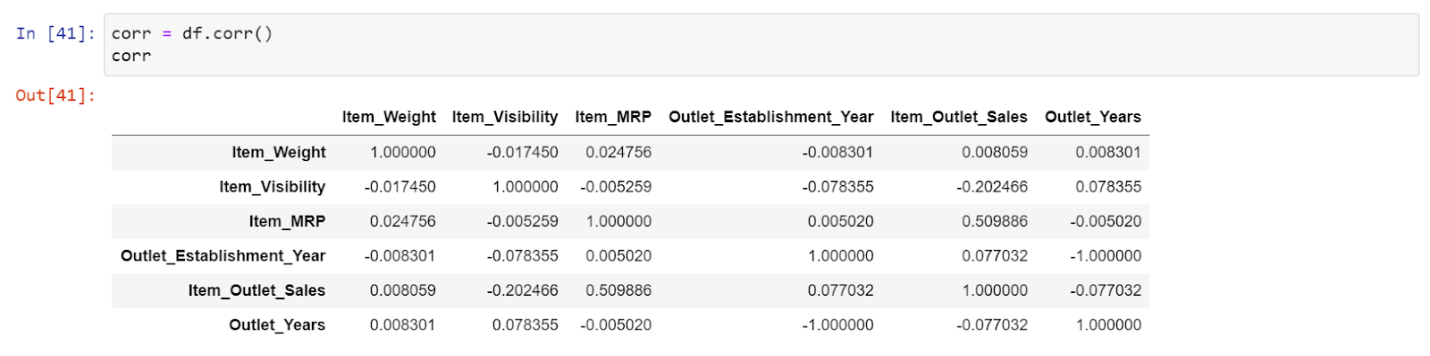
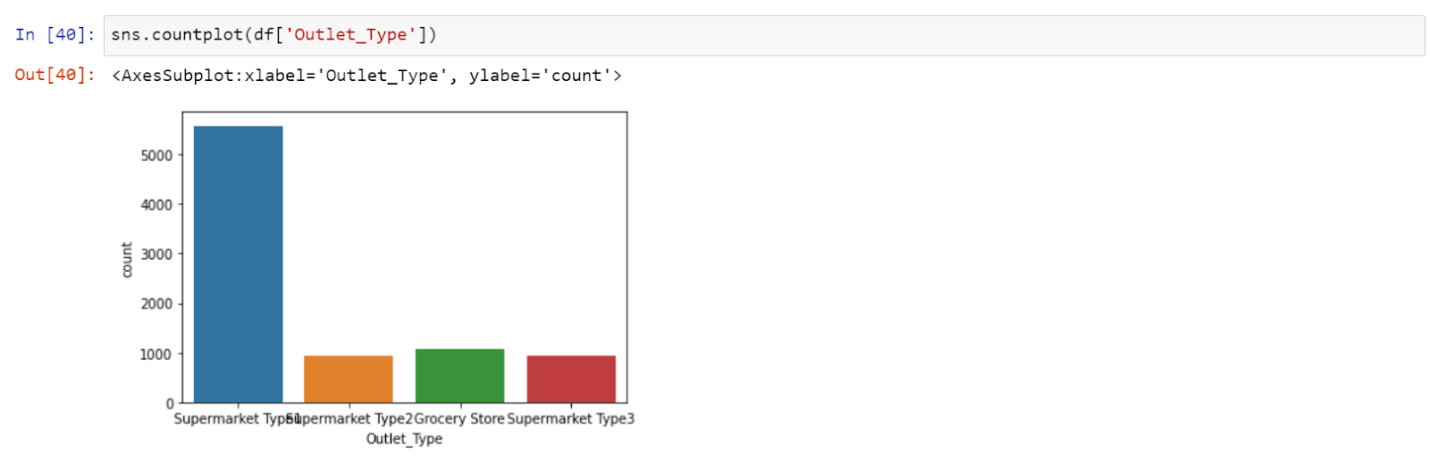
Item type bar graph



Outlet establishment year

Outlet size bar graph

Outlet location type bar graph

Outlet type bar graph

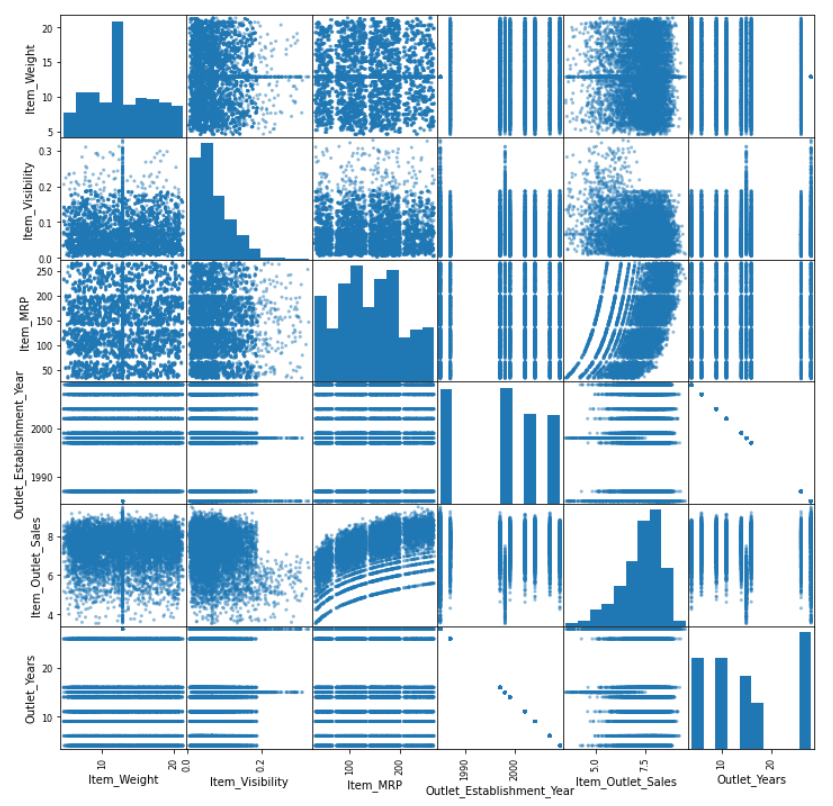
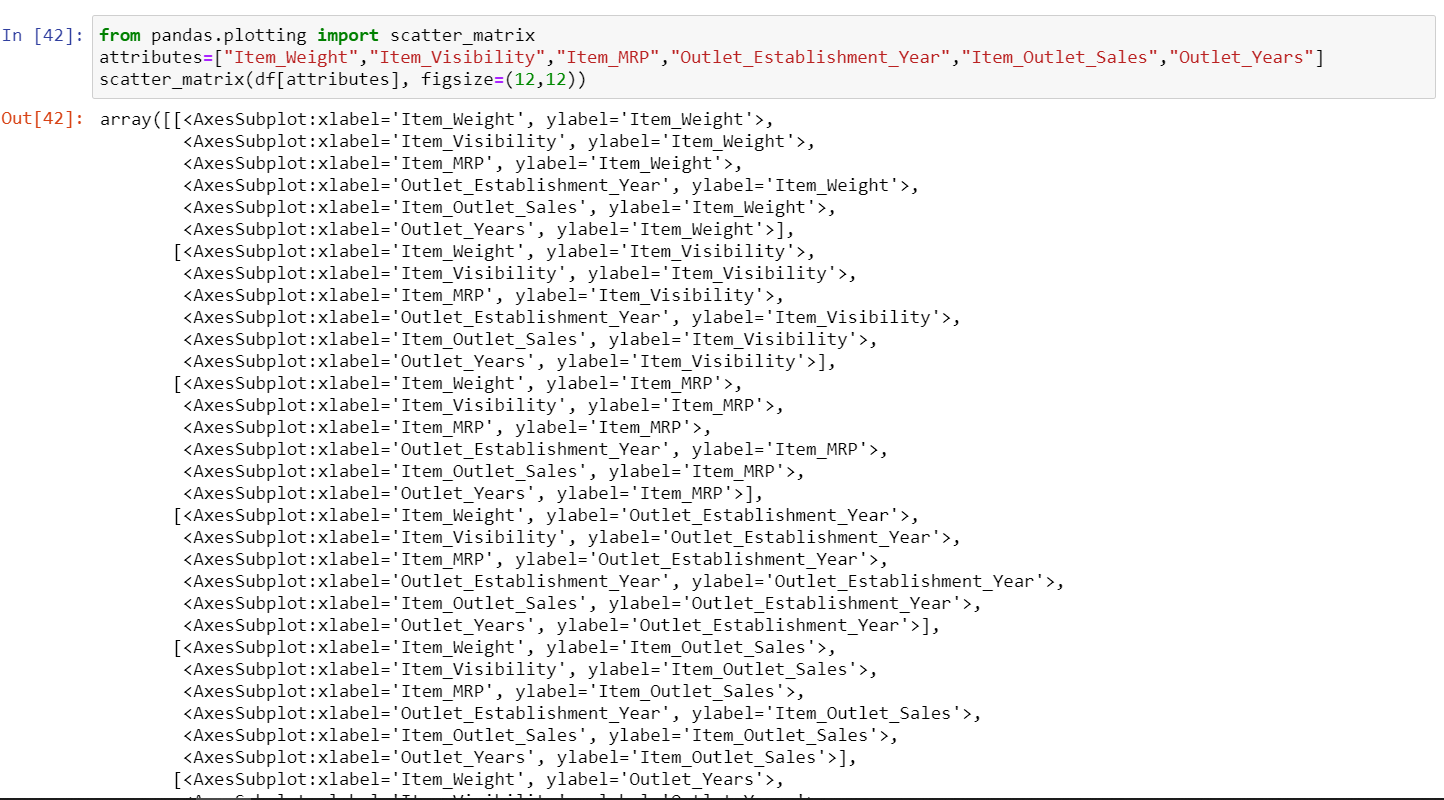
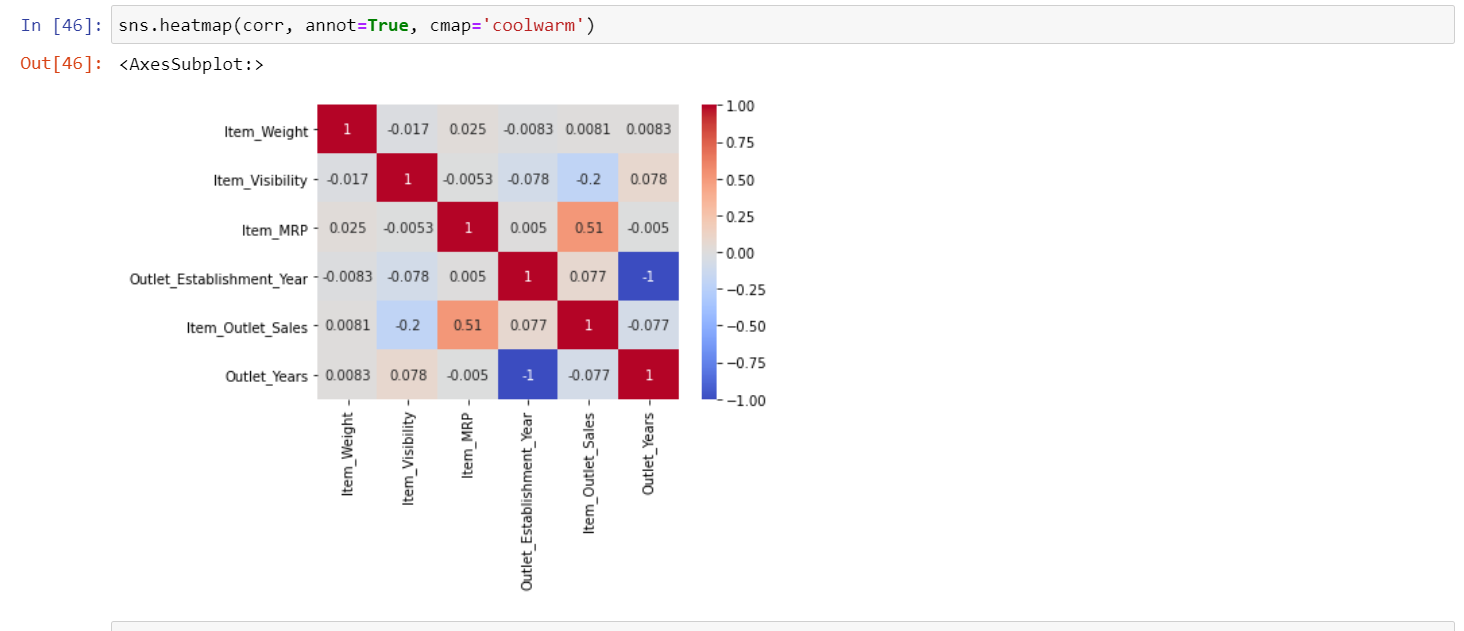
Correlation graphs 

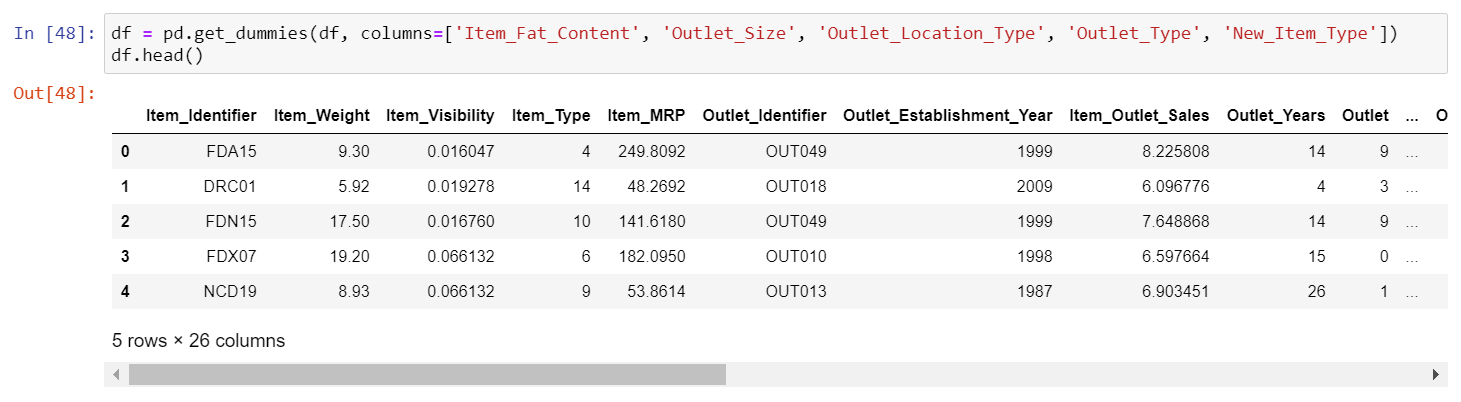
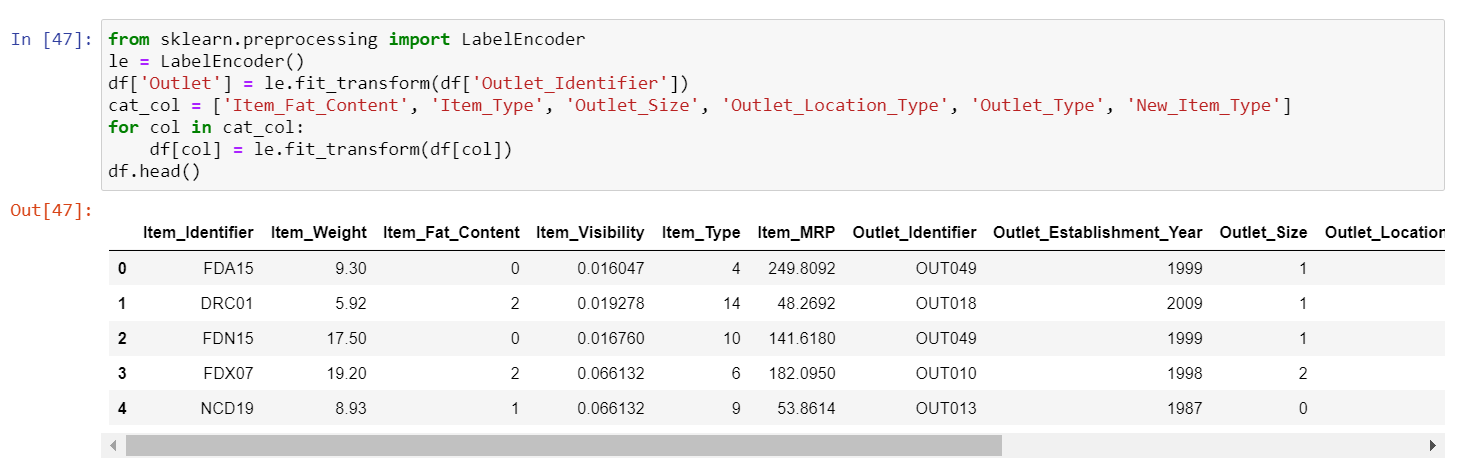
Fig 3

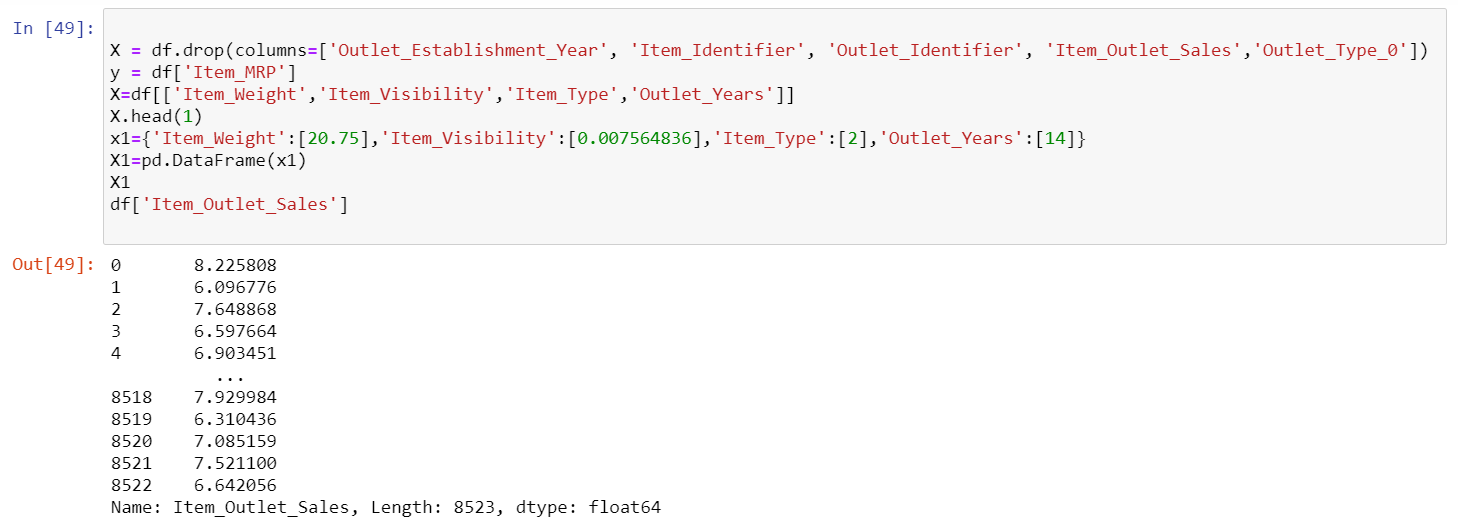




**DATA ANALYSIS:**

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.



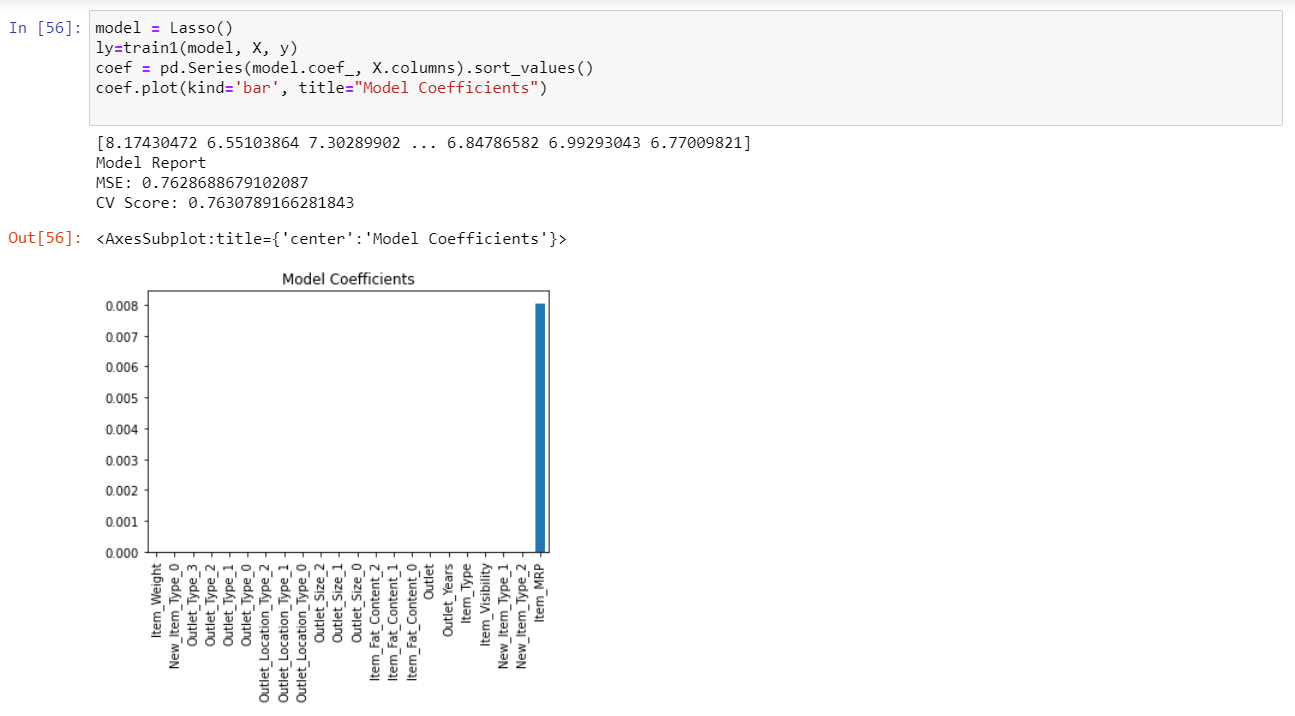


TRAIN MODEL:



**Lasso:**

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso produce encourages simple, spare models. This particular type of regression is well – suited for models showing height levels of multi collinearity or when you want to automate certain parts of model selection, like variable selection or parameter selection. In lasso algorithm, the Item MRP is high so we can have best predictions for item MRP.

fig 6

**Decision Tree:**

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, each representing values for the attribute tested. Leaf node represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. In decision tree algorithm outlet-type-0 and item MRP are having good values compared to other columns.

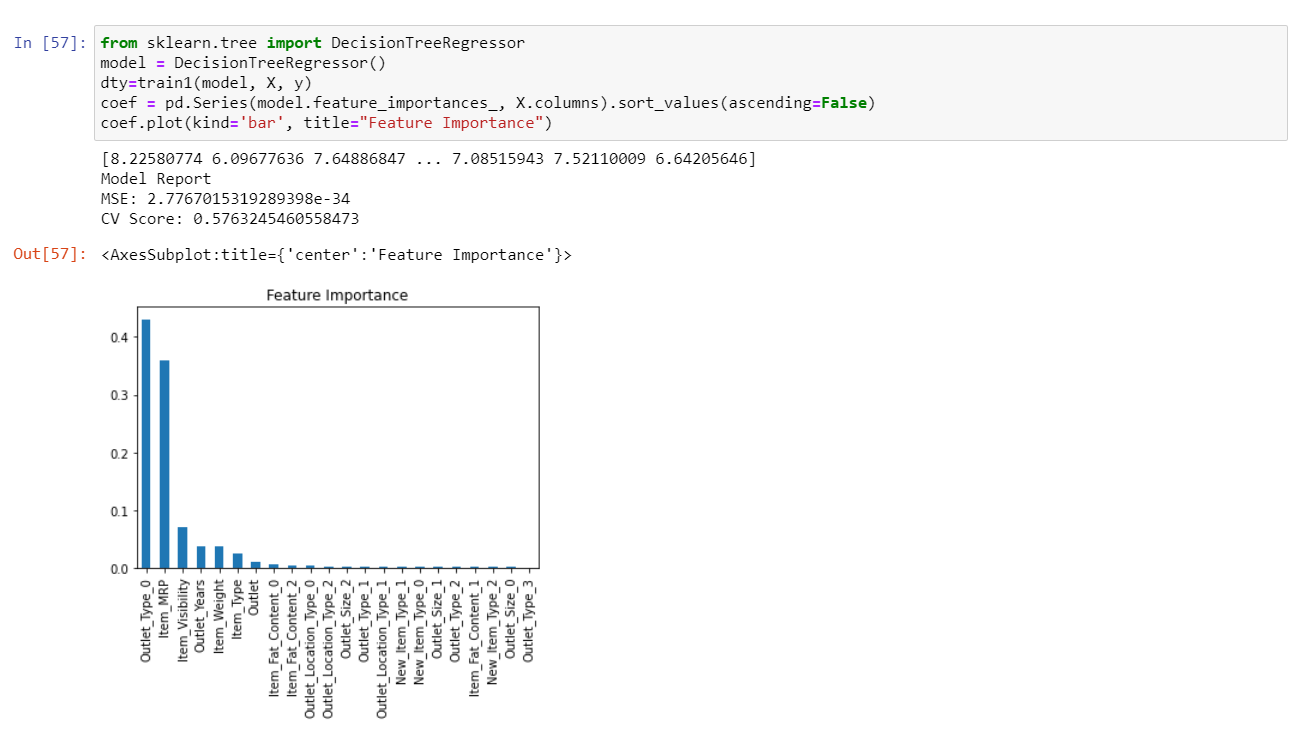


Fig 6

**Random Forest Regression:**

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging. What is bagging you may ask? Bagging, in the Random Forest method, involves training each decision tree on a different data sample where sampling is done with replacement. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees