SAVITRIBAI PHULE PUNE UNIVERSITY A MINI

PROJECT REPORT ON

Big Mart Sales Prediction

SUBMITTED TOWARDS THE
PARTIAL FULFILLMENT OF THE REQUIREMENTS OF

BACHELOR OF ENGINEERING (Computer Engineering)

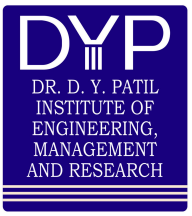
BY

Prasann Shimpi	71922822M
Sanket Halake	71922556G
Shivam Dharmshetti	71922510J
Shashank Singh	71922816G

Under The Guidance of

Prof. Amol Dhakne
Academic Year 2020-21

DEPARTMENT OF COMPUTER ENGINEERING Dr. D. Y. Patil Institute of Engineering, Management & Research Akurdi, Pune.



Dr. D. Y. Patil Institute of Engineering, Management & Research

DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that the Project Entitled

Big Mart Sales Prediction

Submitted by

Prasann Shimpi	71922822M
Sanket Halake	71922556G
Shivam Dharmshetti	71922510J
Shashank Singh	71922816G

is a bonafide work carried out by Students under the supervision of **Prof. Batch Incharge Name** and it is submitted towards the partial fulfillment of the requirement of Bachelor of Engineering (Computer Engineering).

Dr. Amol Dhakne
Internal Guide
Dept. of Computer Engg.

Prof. P.P. Shevatekar
H.O.D
Dept. of Computer Engg.

Dr. A. V. Patil
Principal
Dr. D. Y. Patil Institute of Engineering, Management & Research

Sign of Internal Examiner

Sign of External Examiner

PROJECT APPROVAL SHEET

A Project Title

Is successfully completed by

Prasann Shimpi	71922822M
Sanket Halake	71922556G
Shivam Dharmshetti	71922510J
Shashank Singh	71922816G

at

DEPARTMENT OF COMPUTER ENGINEERING

(Dr. D. Y. Patil Institute of Engineering, Management & Research) SAVITRIBAI PHULE

PUNE UNIVERSITY,PUNE ACADEMIC YEAR 2020-2021

Dr. Amol Dhakne
Internal Guide
Dept. of Computer Engg.

Prof. P.P.Shevatekar
H.O.D
Dept. of Computer Engg.

Abstract

Today, malls and major marts track sales data for all items to predict future customer demand and update inventory management. These databases basically contain a large amount of customer data and individual item attributes in the data warehouse. It also detects anomalies and common patterns by mining the data store from the data warehouse. The data obtained can be used to predict future sales volumes using a variety of machine learning techniques for retailers such as Big Mart. In this paper, we proposed a forecasting model that uses the Xgboost method to predict sales for a company like Big Mart, and found that the model performed better than existing models. It also details the comparative analysis of key performance indicators with other models.

Acknowledgments

It gives us great pleasure in presenting the preliminary project report on 'Project Title'.

I would like to take this opportunity to thank my internal guide **Dr. Amol Dhakne** for giving me all the help and guidance I needed. I am really grateful to them for their kind support. Their valuable suggestions were very helpful.

I am also grateful to Prof. P.P. Shevtekar, Head of Computer Engineer- ing Department, Dr. D. Y. Patil Institute of Engineering, Management & Research for his indispensable support, suggestions.

In the end our special thanks to **Dr. Amol Dhakne** for providing various resources such as laboratory with all needed software platforms, continuous Internet connection, for Our Project.

Prasann Shimpi	71922822M
Sanket Halake	71922556G
Shivam Dharmshetti	71922510J
Shashank Singh	71922816G
	(B.E. Computer Engg.)

Contents

1.

1.1. Abstract

Today, malls and major marts track sales data for all items to predict future customer demand and update inventory management. These databases basically contain a large amount of customer data and individual item attributes in the data warehouse. It also detects anomalies and common patterns by mining the data store from the data warehouse. The data obtained can be used to predict future sales volumes using a variety of machine learning techniques for retailers such as Big Mart. In this paper, we proposed a forecasting model that uses the Xgboost method to predict sales for a company like Big Mart, and found that the model performed better than existing models. It also details the comparative analysis of key performance indicators with other models.

1.2. Problem Statement:

Use data mining and warehousing, data analysis and machine learning concepts to build a model capable of making accurate predictions related to the sales of a particular item based on various categories.

1.3. Goals and Objectives

To build a model capable of making accurate predictions about the sales of an item based on ten different attributes of the said item.

2. Technical Words

2.1. Area of Project

The project helps to determine the sales of a particular item based on its sales history. The project implements concepts of data mining and warehousing, data analysis and machine learning.

2.2. Technical Keywords

Machine Learning, Sales Forecasting, Random Forest, Regression

3. Introduction

3.1. Project Idea

Due to the rapid growth of global shopping malls and online shopping, competition between various shopping malls and large markets is intensifying every day. All malls or marts seek to offer personalized short-term offers to attract more customers by day. This allows you to predict how much each item will be sold in your organization's inventory management, logistics, and transportation services. The learning algorithm is very sophisticated and provides a method for forecasting or forecasting the future sales demand of a company. It also helps overcome the low availability of computers and storage systems. This article discusses the problem of forecasting or predicting items in terms of future customer demand in different locations and different large markets for products based on previous datasets. Use various machine learning algorithms such as linear regression analysis and random forest to forecast or predict sales volume. Forecasting plays an important role in every shopping complex, as good sales are the life of every business. Always better forecasts help you develop and improve your market business strategy and help you improve your knowledge of the market.

3.2. Motivation of the Project

During these tough times due to the pandemic, the sales of different commodities have varied drastically when compared in a larger time frame. As lockdowns are being lifted and everyone has started to resume their normal routine at a steady pace, the previous sales data is crucial for managing the demand supply requirements of a particular commodity. Thus, a model capable of predicting such sales without human effort whilst saving time is very beneficial for mankind. As Big Mart Sales dataset was available on Kaggle, we have used is to predict sales based on categories like type of item, outlet size, location, weight, visibility, etc. Using this as a initial model, a few more improvements and then some, it will surely help in regulating the demand and supply cycle.

4. Problem Scope

4.1. Applications

In today's digitally connected world, every shopping center wants to know in advance customer requirements to avoid a shortage of items for sale throughout the year. Every day, businesses and shopping malls more accurately forecast product sales demand and user demand. Extensive research in this area is being conducted at the enterprise level to create accurate sales forecasts. Big Marts wants a more accurate forecasting algorithm so that the company doesn't suffer losses because the company's profits are directly proportional to the exact forecast.

4.2. Hardware Resources

8 GB RAM, i5 10^{th} Gen Processor Laptop/PC.

4.3. Software Resources

Python, Anaconda, Jupyter Notebook

5. Project Implementation

5.1. Introduction

Due to the rapid growth of global shopping malls and online shopping, competition between various shopping malls and large markets is intensifying every day. All malls or marts seek to offer personalized short-term offers to attract more customers by day. This allows you to predict how much each item will be sold in your organization's inventory management, logistics, and transportation services. The learning algorithm is very sophisticated and provides a method for forecasting or forecasting the future sales demand of a company. It also helps overcome the low availability of computers and storage systems. This article discusses the problem of forecasting or predicting items in terms of future customer demand in different locations and different large markets for products based on previous datasets. Use various machine learning algorithms such as linear regression analysis and random forest to forecast or predict sales volume. Forecasting plays an important role in every shopping complex, as good sales are the life of every business. Always better forecasts help you develop and improve your market business strategy and help you improve your knowledge of the market. A standard forecasting survey thoroughly analyzes previously encountered situations and situations, draws conclusions about customer acquisition, resource shortages, and strengths, and then sets a budget and marketing plan one year ahead. Useful for. In other words, forecasts are based on past available resources.

5.2. Tools and Technologies

5.2.1. Tool

Python - Anaconda - Jupyter Notebook

5.2.2.Technology

Python Libraries

- Scikit-learn
- Pandas
- numpy
- Matplotlib
- Dtate
- klib

5.3. Results

5.3.1.Screenshots

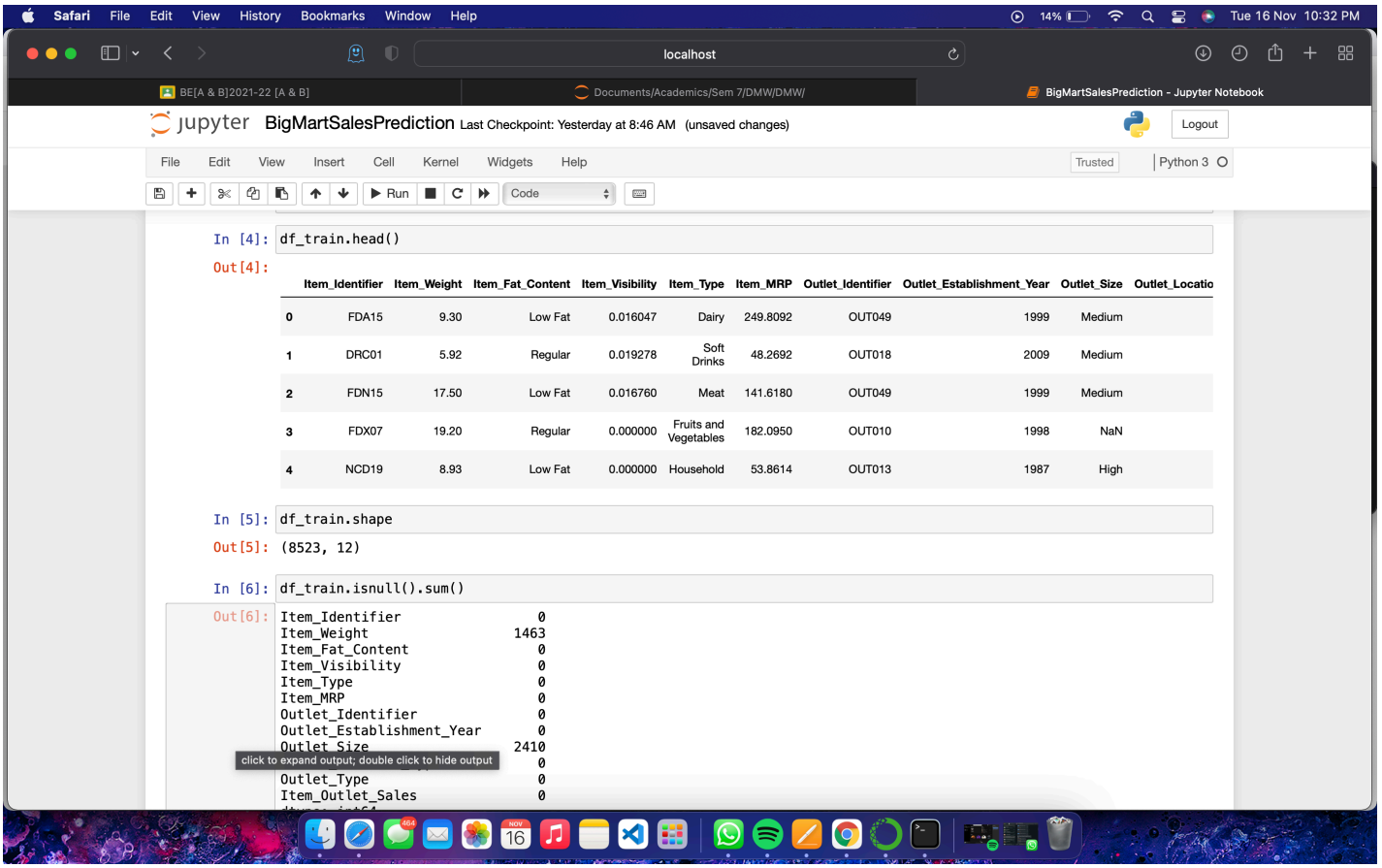
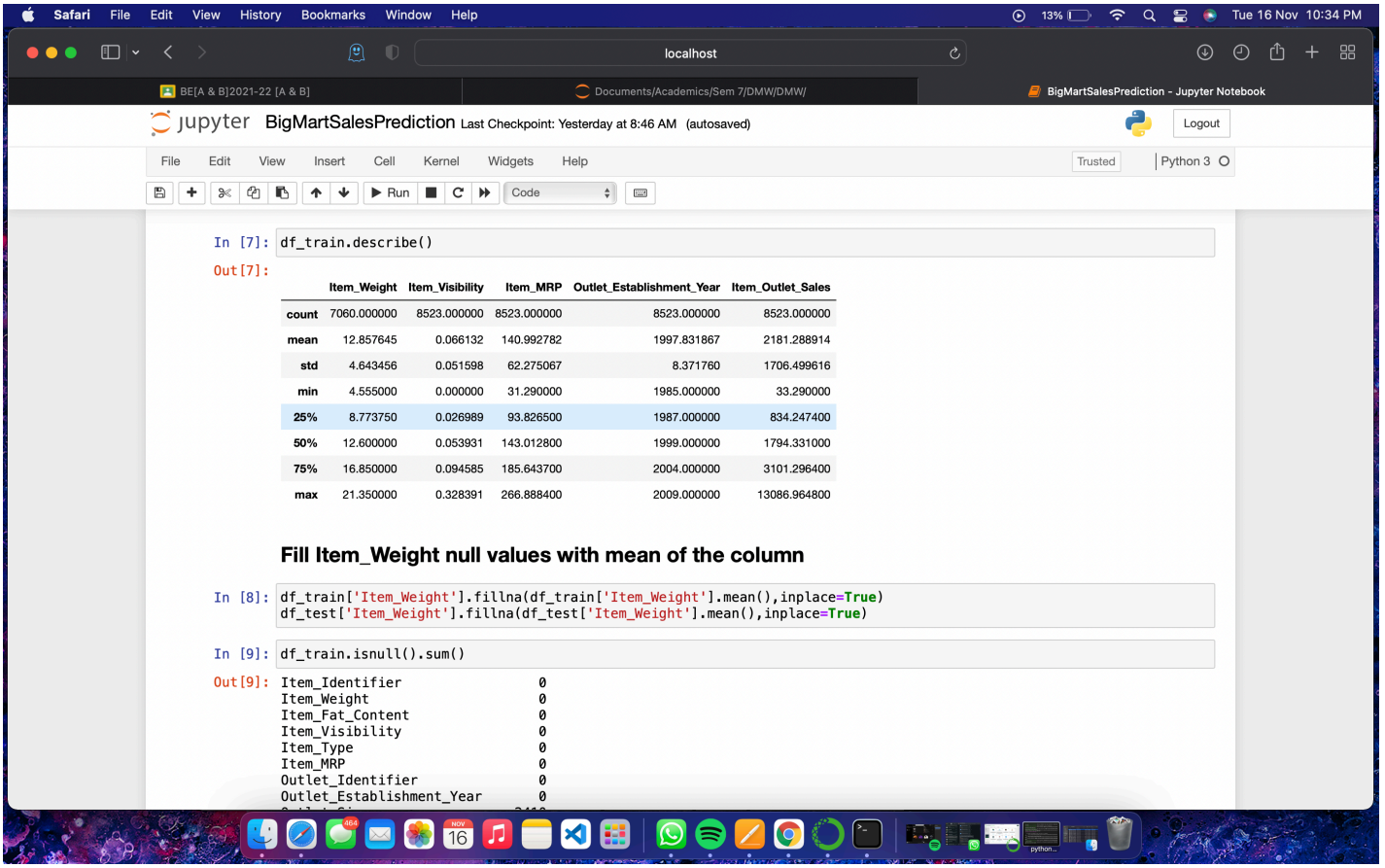


Fig 1: Dataframe attributes

Fig 2 and 3: Eliminating Null Values



jupyter BigMartSalesPrediction Last Checkpoint: Yesterday at 8:46 AM (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Code

In [7]:
```
df_train.describe()
```

Out[7]:

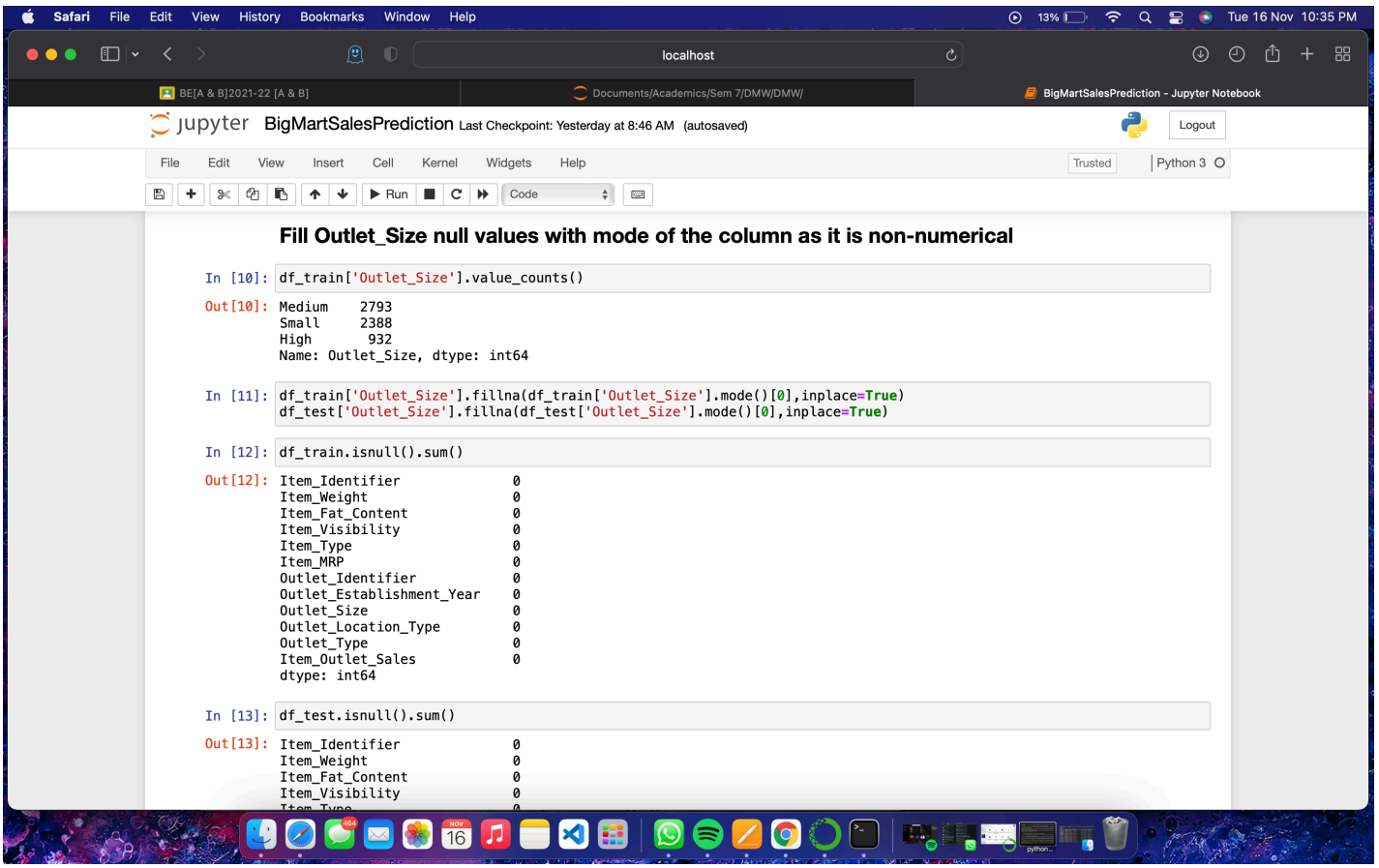
	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Outlet_Sales
count	7060.000000	8523.000000	8523.000000	8523.000000	8523.000000
mean	12.857645	0.066132	140.992782	1997.831867	2181.288914
std	4.643456	0.051598	62.275067	8.371760	1706.499616
min	4.555000	0.000000	31.290000	1985.000000	33.290000
25%	8.773750	0.026989	93.826500	1987.000000	834.247400
50%	12.600000	0.053931	143.012800	1999.000000	1794.331000
75%	16.850000	0.094585	185.643700	2004.000000	3101.296400
max	21.350000	0.328391	266.888400	2009.000000	13086.964800

Fill Item_Weight null values with mean of the column

In [8]:
```
df_train['Item_Weight'].fillna(df_train['Item_Weight'].mean(),inplace=True)
df_test['Item_Weight'].fillna(df_test['Item_Weight'].mean(),inplace=True)
```

In [9]:
```
df_train.isnull().sum()
```

Out[9]:
```
Item_Identifier                 0
Item_Weight                     0
Item_Fat_Content                0
Item_Visibility                 0
Item_Type                       0
Item_MRP                        0
Outlet_Identifier               0
Outlet_Establishment_Year       0
```



jupyter BigMartSalesPrediction Last Checkpoint: Yesterday at 8:46 AM (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Code

Fill Outlet_Size null values with mode of the column as it is non-numerical

In [10]:
```
df_train['Outlet_Size'].value_counts()
```

Out[10]:
```
Medium    2793
Small     2388
High       932
Name: Outlet_Size, dtype: int64
```

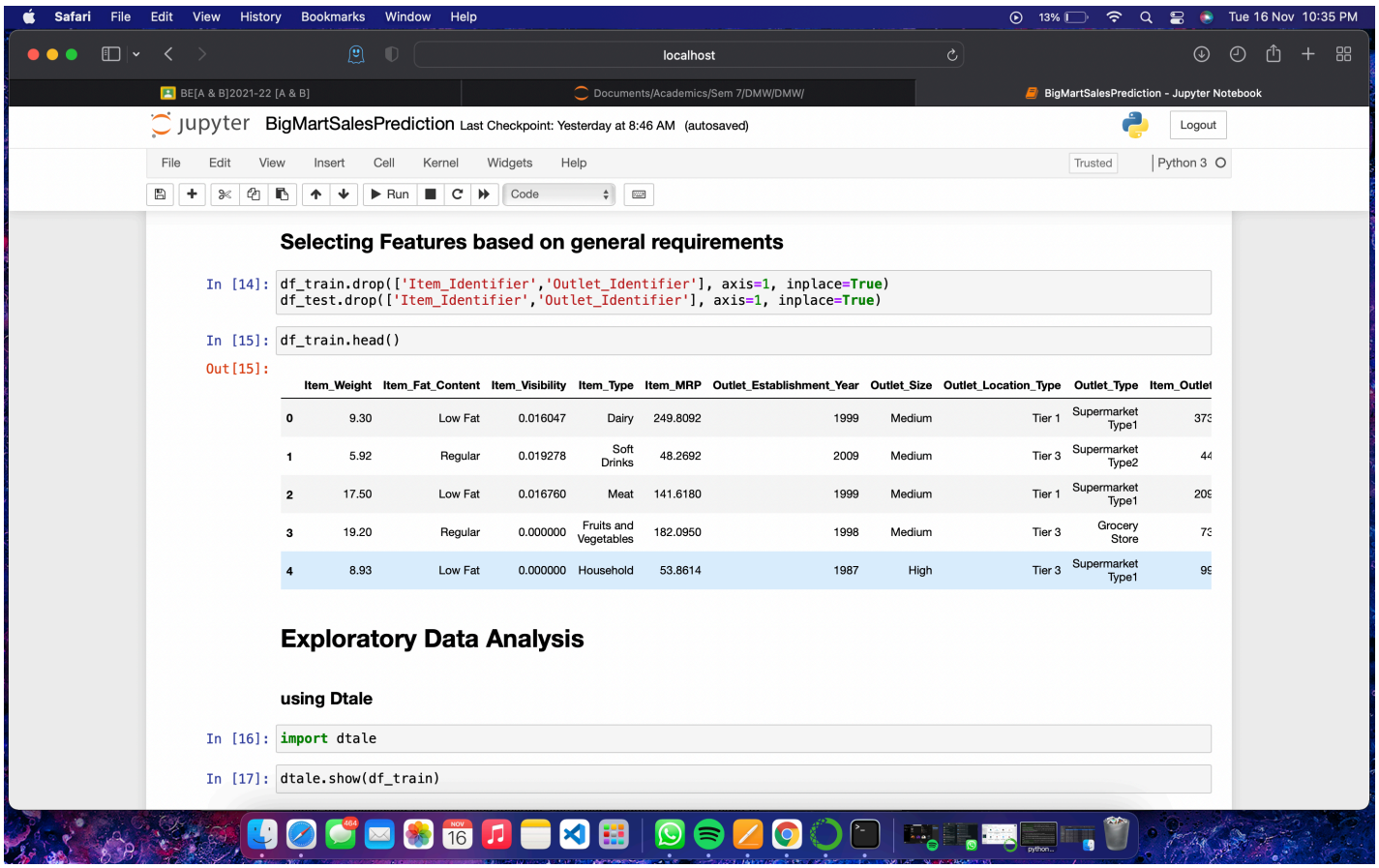
In [11]:
```
df_train['Outlet_Size'].fillna(df_train['Outlet_Size'].mode()[0],inplace=True)
df_test['Outlet_Size'].fillna(df_test['Outlet_Size'].mode()[0],inplace=True)
```

In [12]:
```
df_train.isnull().sum()
```

Out[12]:
```
Item_Identifier              0
Item_Weight                  0
Item_Fat_Content              0
Item_Visibility               0
Item_Type                     0
Item_MRP                      0
Outlet_Identifier             0
Outlet_Establishment_Year     0
Outlet_Size                   0
Outlet_Location_Type          0
Outlet_Type                   0
Item_Outlet_Sales             0
dtype: int64
```

In [13]:
```
df_test.isnull().sum()
```

Out[13]:
```
Item_Identifier              0
Item_Weight                  0
Item_Fat_Content              0
Item_Visibility               0
```



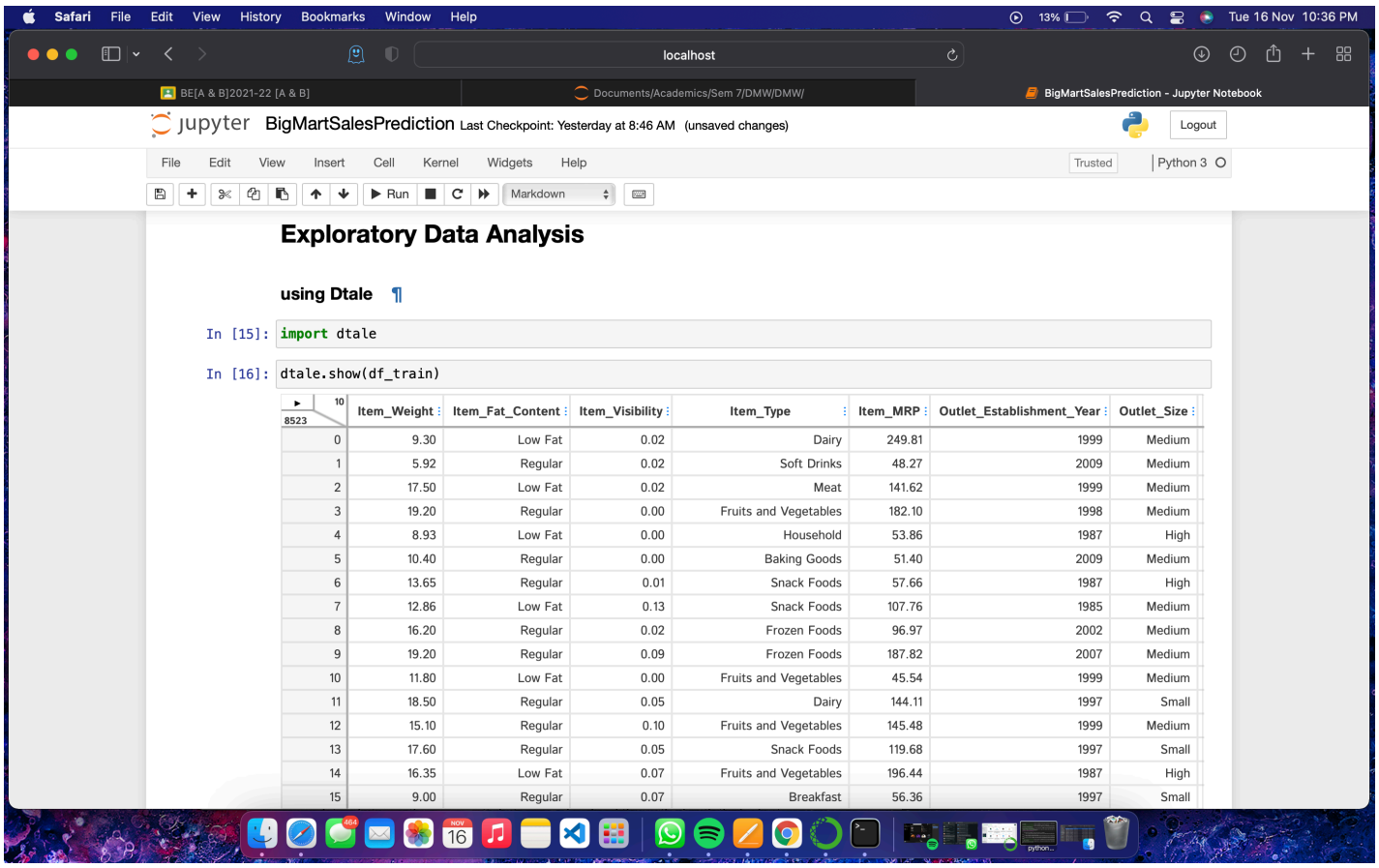
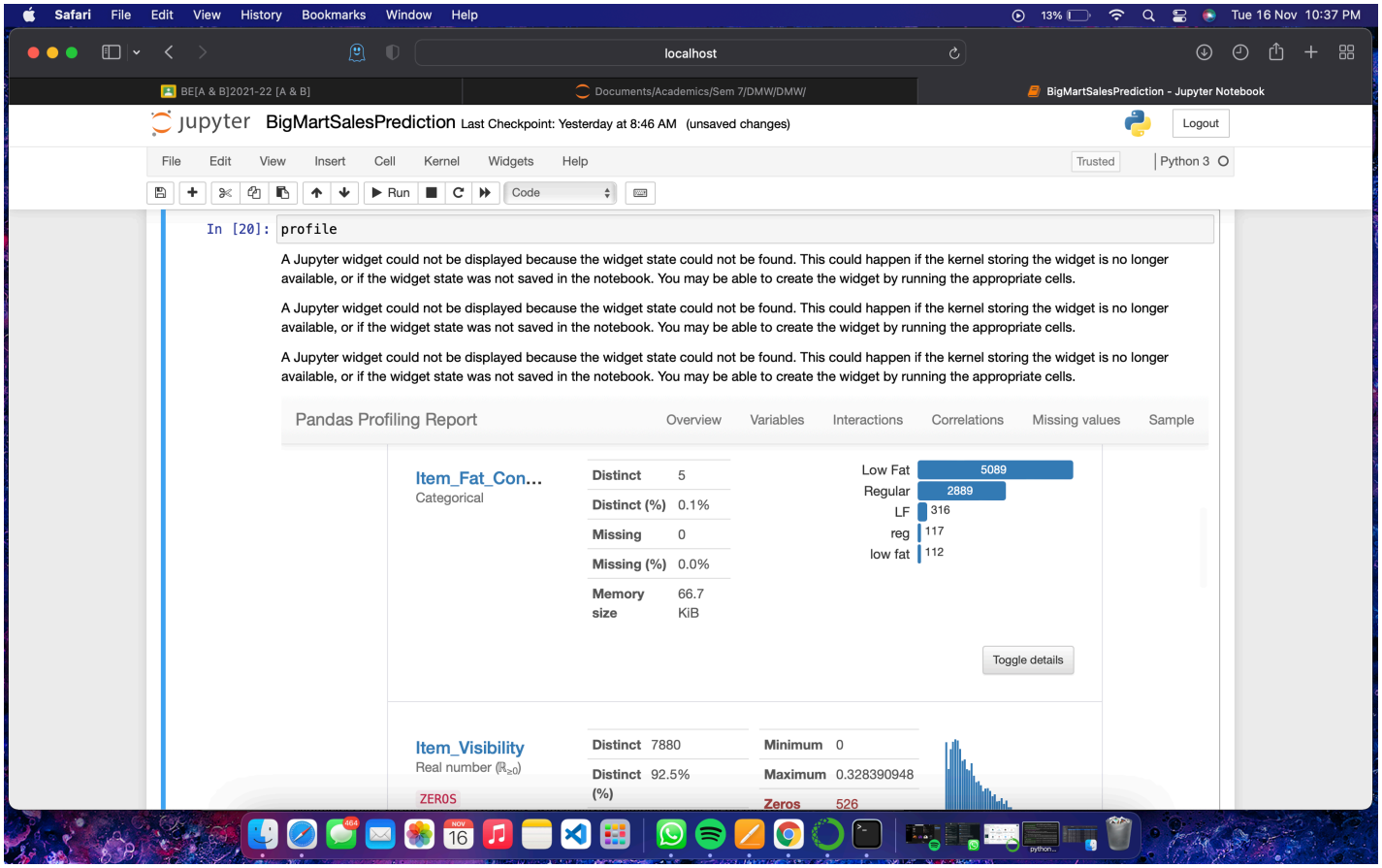


Fig 4 and 5: Data exploration and Analysis



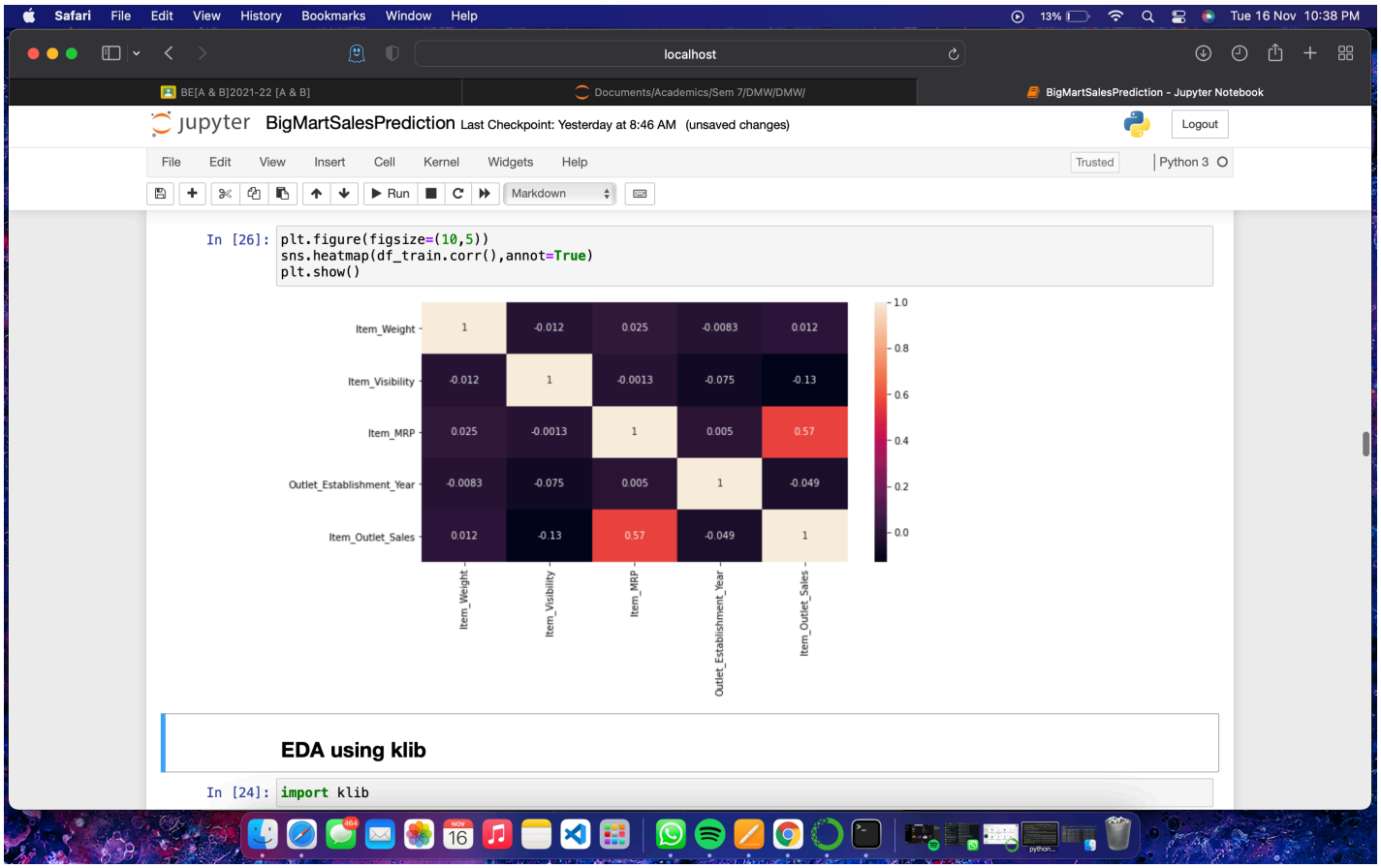
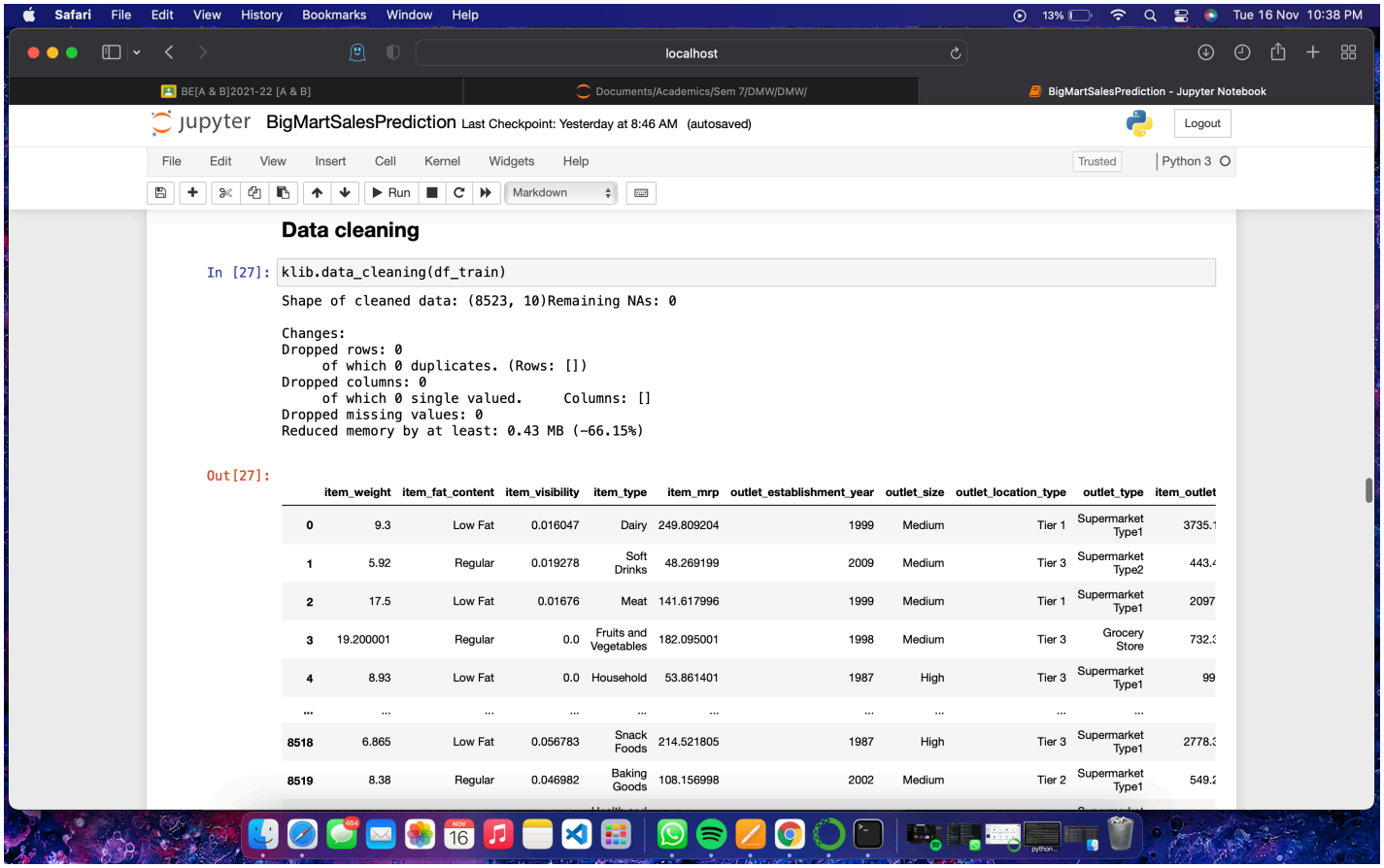
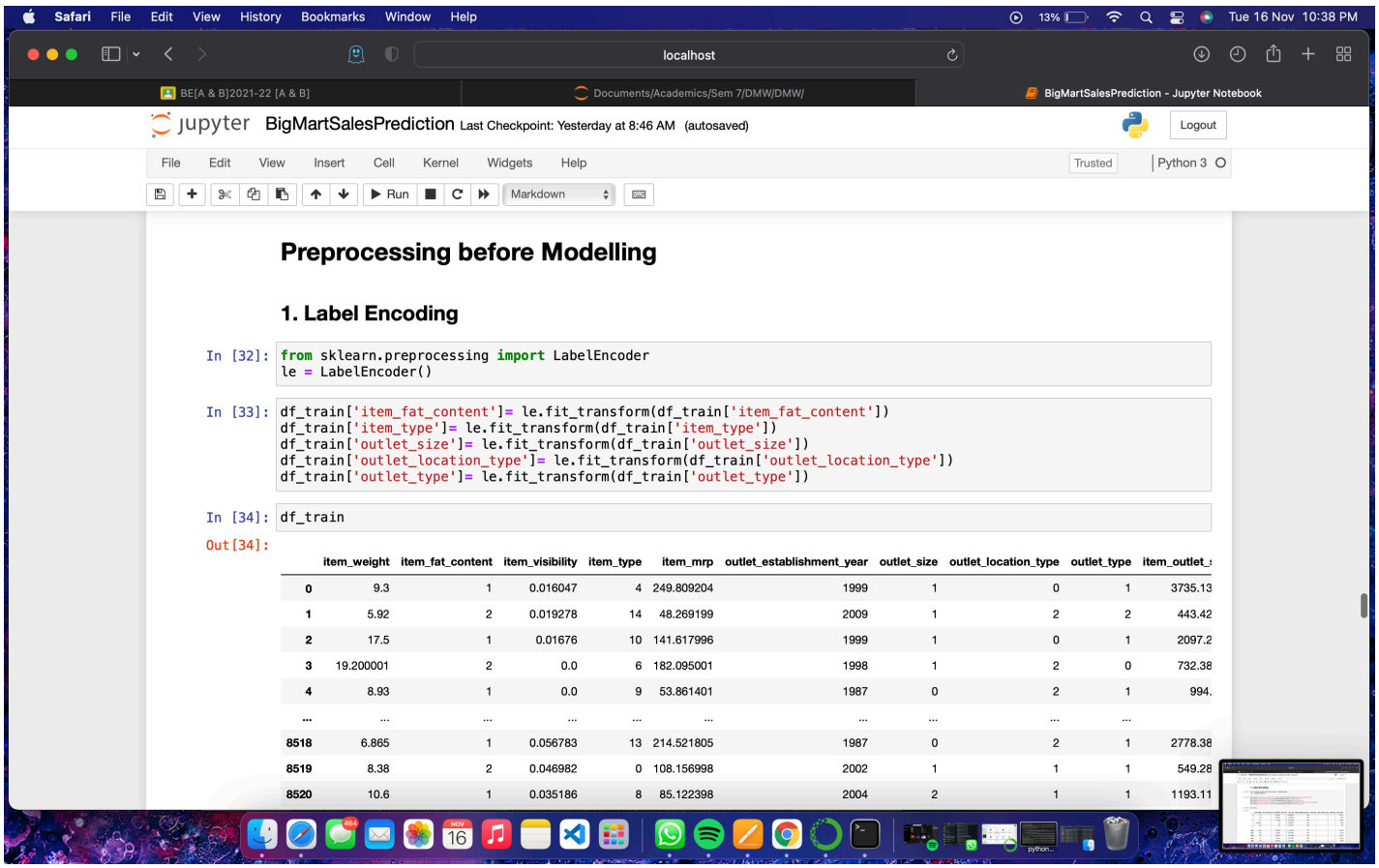


Fig 6 and 7: Data Visualisation

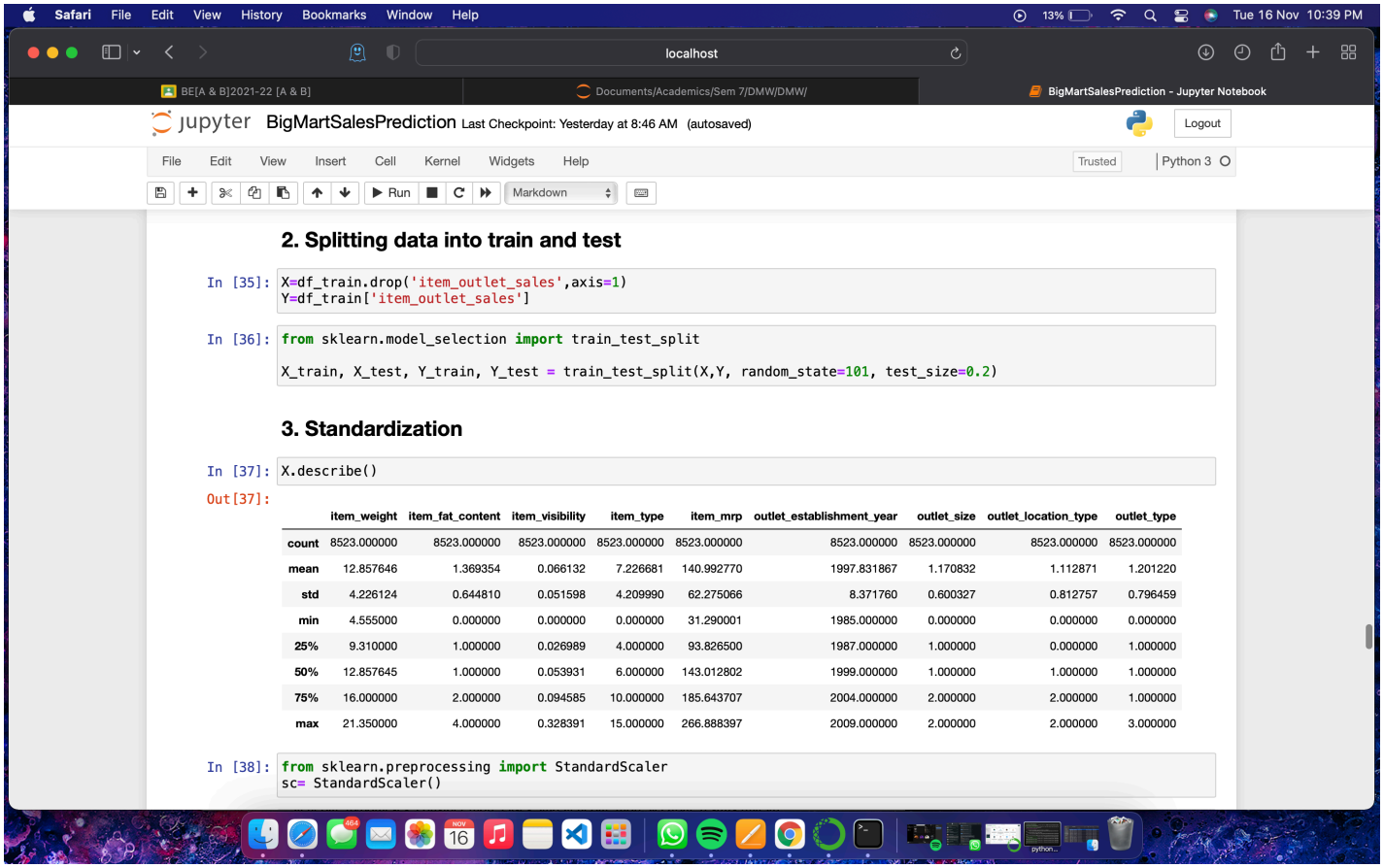


	item_weight	item_fat_content	item_visibility	item_type	item_mrp	outlet_establishment_year	outlet_size	outlet_location_type	outlet_type	item_outlet
0	9.3	Low Fat	0.016047	Dairy	249.809204	1999	Medium	Tier 1	Supermarket Type1	3735.1
1	5.92	Regular	0.019278	Soft Drinks	48.269199	2009	Medium	Tier 3	Supermarket Type2	443.4
2	17.5	Low Fat	0.01676	Meat	141.617996	1999	Medium	Tier 1	Supermarket Type1	2097
3	19.200001	Regular	0.0	Fruits and Vegetables	182.095001	1998	Medium	Tier 3	Grocery Store	732.3
4	8.93	Low Fat	0.0	Household	53.861401	1987	High	Tier 3	Supermarket Type1	99
...	...	...	...	...	...	...	...	...	...	...
8518	6.865	Low Fat	0.056783	Snack Foods	214.521805	1987	High	Tier 3	Supermarket Type1	2778.3
8519	8.38	Regular	0.046982	Baking Goods	108.156998	2002	Medium	Tier 2	Supermarket Type1	549.2



	item_weight	item_fat_content	item_visibility	item_type	item_mrp	outlet_establishment_year	outlet_size	outlet_location_type	outlet_type	item_outlet_
0	9.3	1	0.016047	4	249.809204	1999	1	0	1	3735.13
1	5.92	2	0.019278	14	48.269199	2009	1	2	2	443.42
2	17.5	1	0.01676	10	141.617996	1999	1	0	1	2097.2
3	19.200001	2	0.0	6	182.095001	1998	1	2	0	732.38
4	8.93	1	0.0	9	53.861401	1987	0	2	1	994.
...	...	...	...	...	...	...	...	...	...	...
8518	6.865	1	0.056783	13	214.521805	1987	0	2	1	2778.38
8519	8.38	2	0.046982	0	108.156998	2002	1	1	1	549.28
8520	10.6	1	0.035186	8	85.122398	2004	2	1	1	1193.11

Fig 8 and 9 : Data Cleaning and Label Encoding



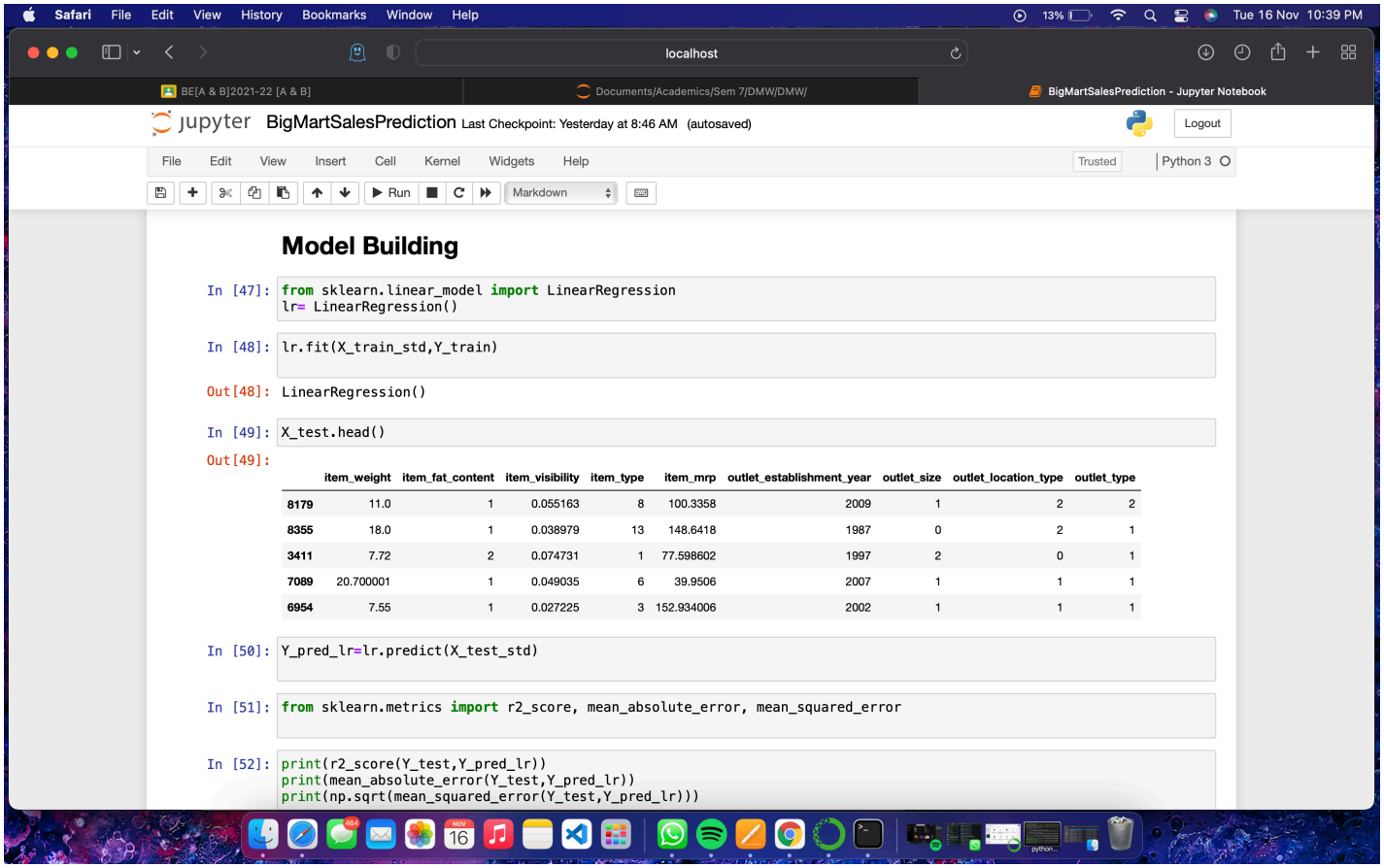
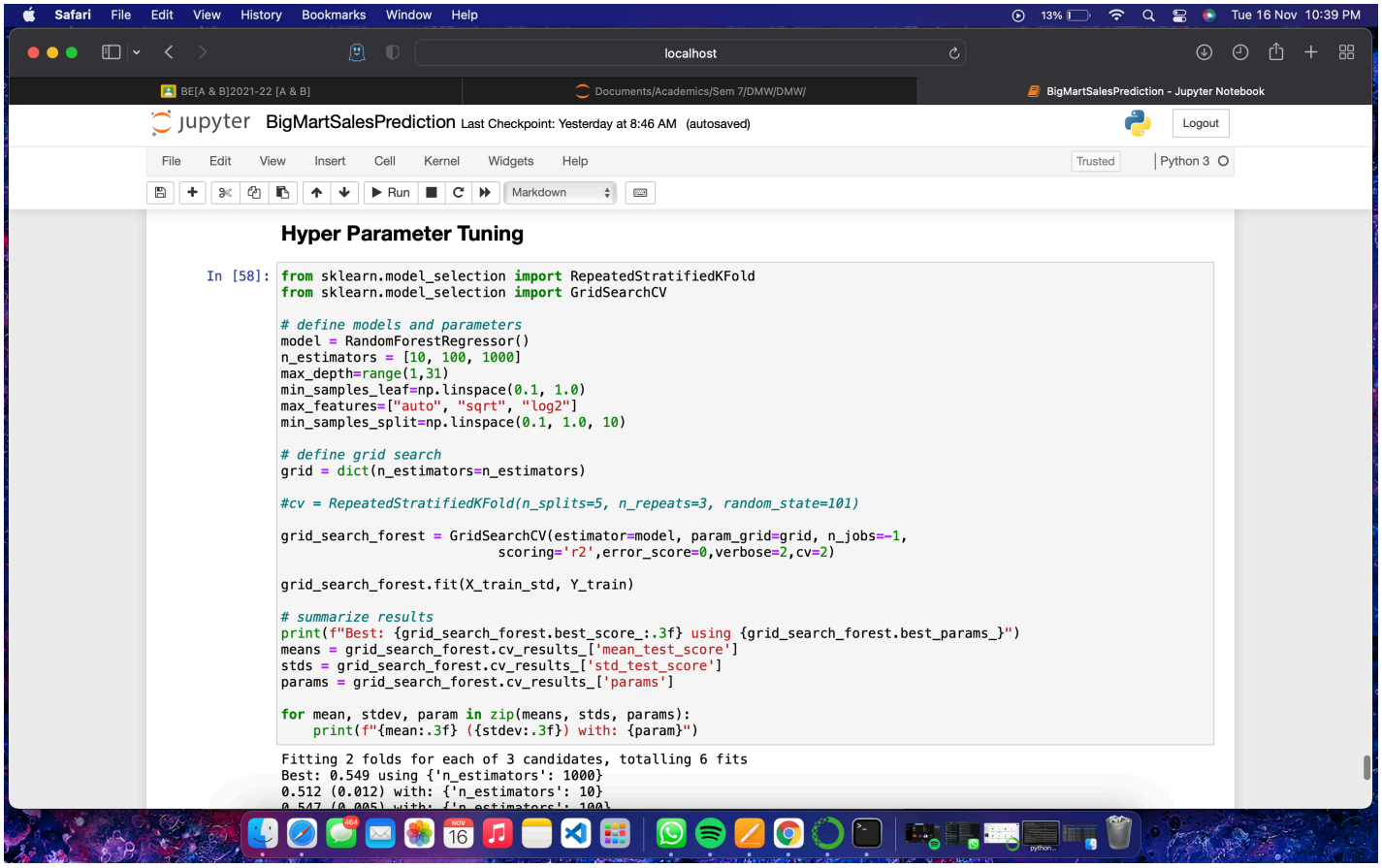


Fig 10 and 11: Data Preparation and Model Building



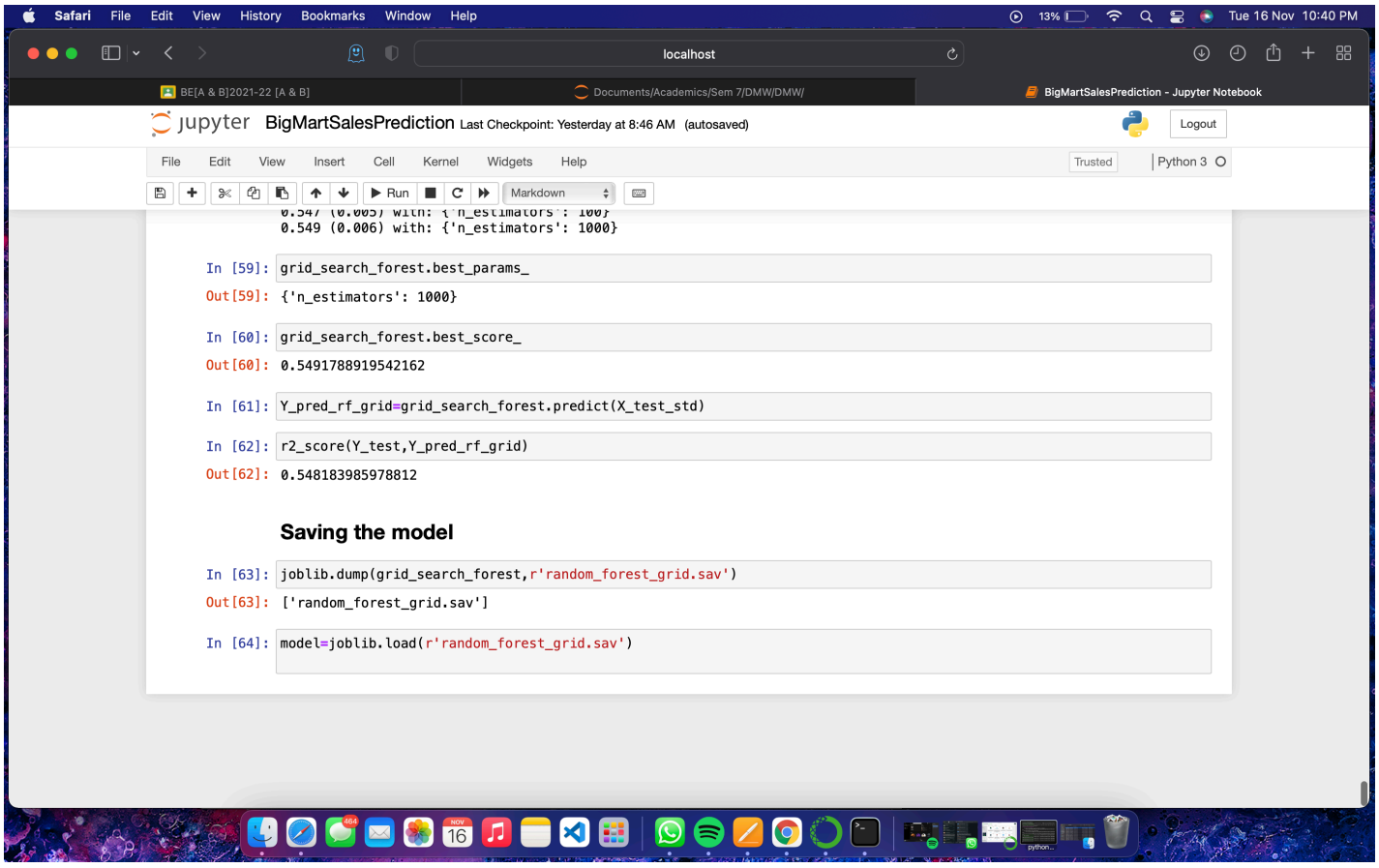
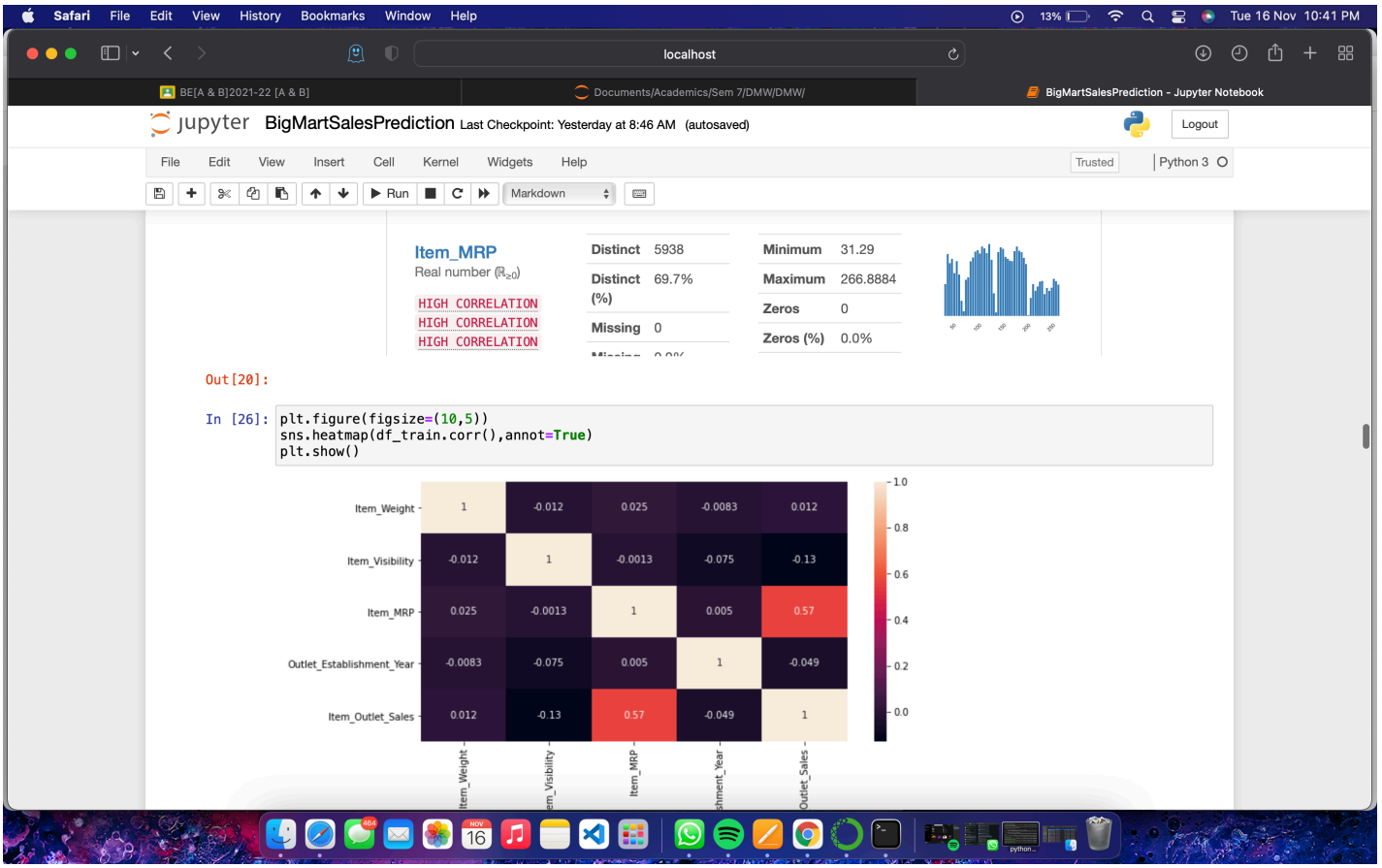
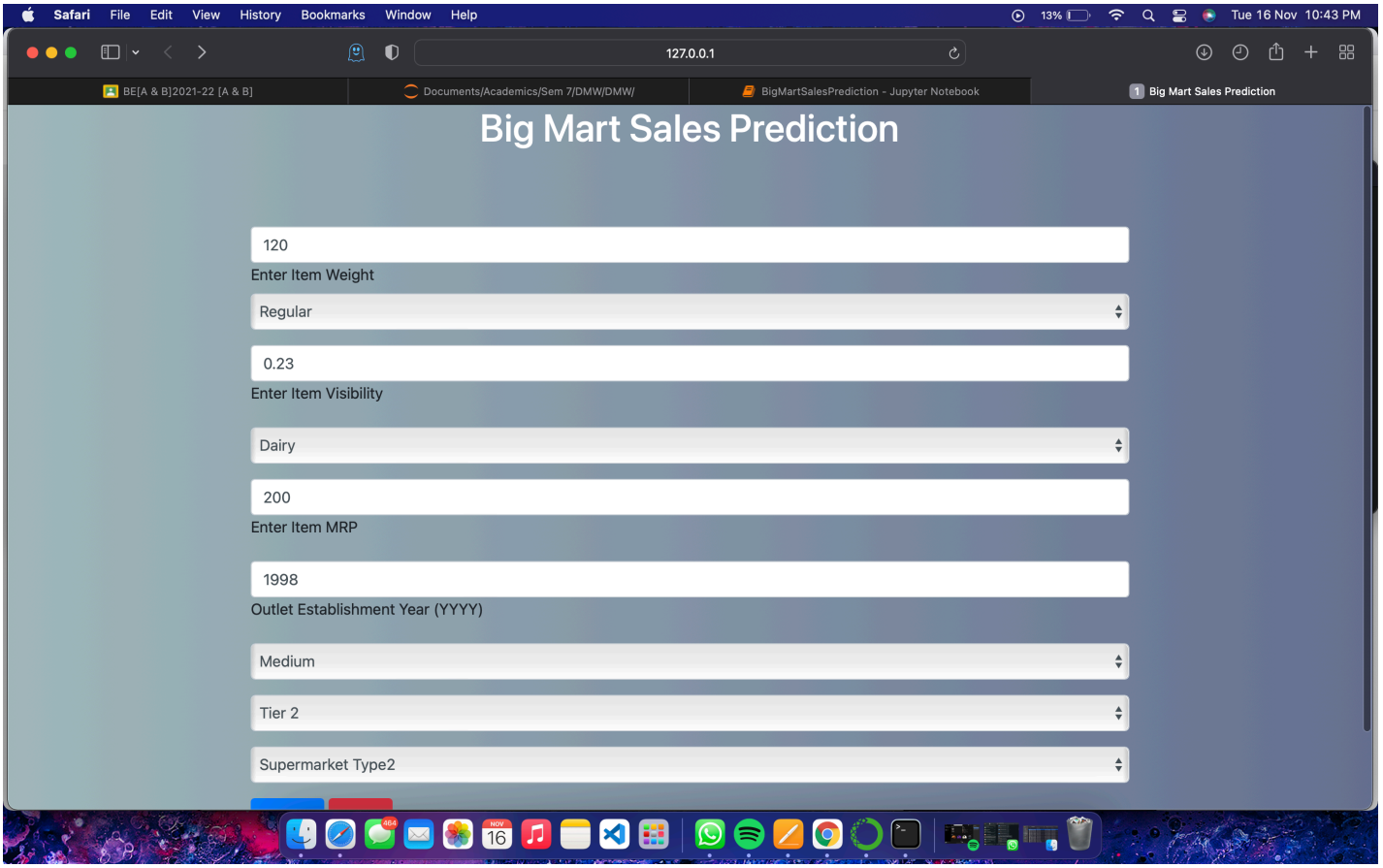


Fig 12 and 13: Hyperparameter tuning and Model Dump

5.3.2.Outputs


Item_MRP
Real number (ℝ≥0)
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
Distinct 5938
Distinct (%) 69.7%
Missing 0
Minimum 31.29
Maximum 266.8884
Zeros 0
Zeros (%) 0.0%
Out[20]:
In [26]: plt.figure(figsize=(10,5))
sns.heatmap(df_train.corr(),annot=True)
plt.show()
Item_Weight
Item_Visibility
Item_MRP
Outlet_Establishment_Year
Item_Outlet_Sales


Big Mart Sales Prediction
120
Enter Item Weight
Regular
0.23
Enter Item Visibility
Dairy
200
Enter Item MRP
1998
Outlet Establishment Year (YYYY)
Medium
Tier 2
Supermarket Type2

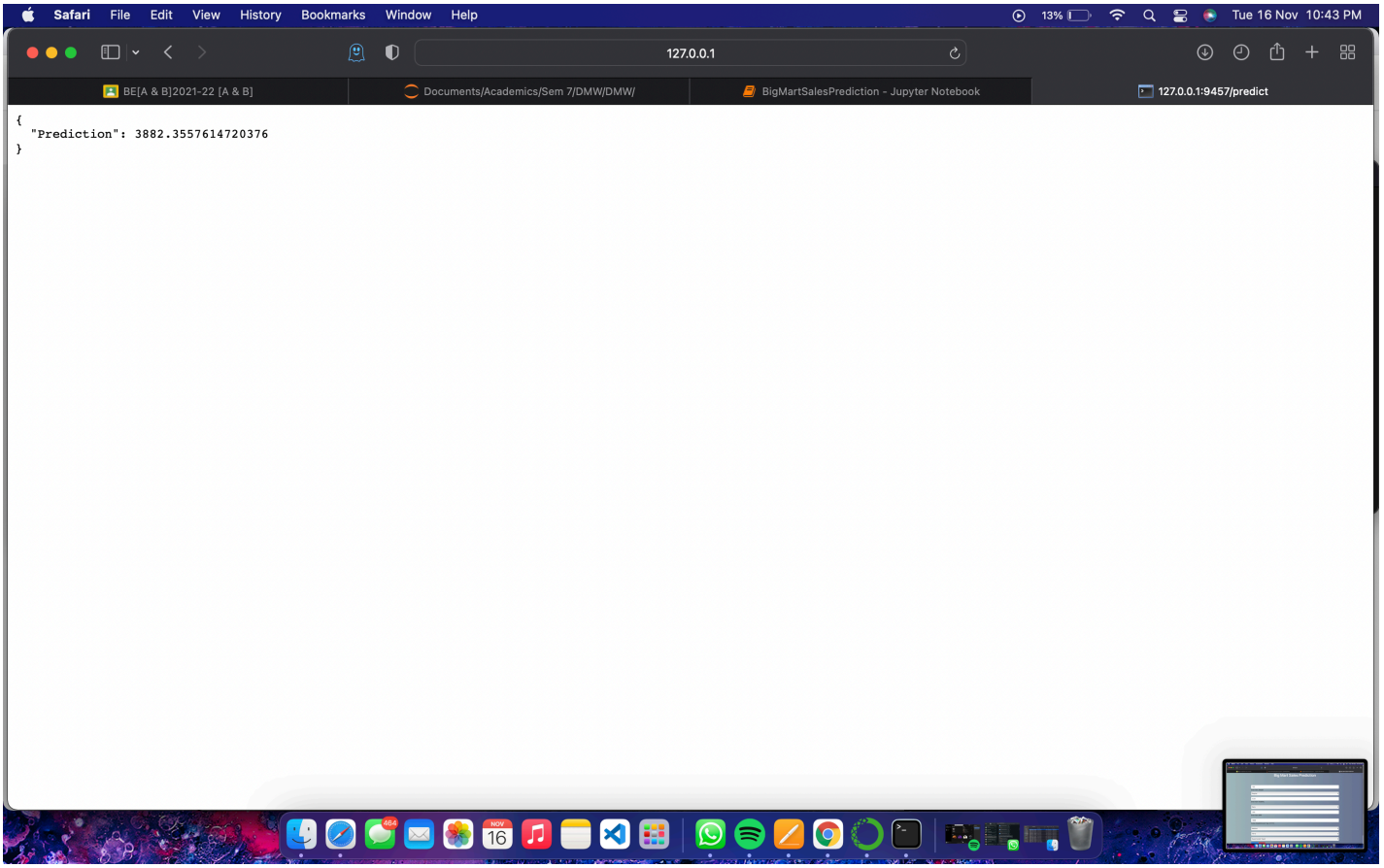


Fig 14, 15 and 16: Correlation Heatmap, Home Page and Sales Prediction

6. Conclusions

In this project, we have explained the basics of machine learning and related data processing and modeling algorithms, and then how to apply them to forecasting tasks at Big Mart Malls in different locations. When implemented, forecast results show the correlation between the various attributes considered and the method with the highest sales for a particular medium-sized location, and other shopping locations need to improve sales according to a similar pattern. Accuracy, which plays an important role in predictive-based systems, can improve significantly as the number of parameters used increases. Checking the functionality of the partial model can also improve the productivity of the system. Projects can be further collaborated using built-in intelligence on web-based applications or any device supported by the Internet of Things (IoT). Various stakeholders working with sales information can also provide more input to help generate hypotheses, consider more cases, and generate more accurate results that are closer to the actual situation. Combined with effective data mining methods and characteristics, traditional tools can have a higher and more positive effect on the overall development of a company's overall tasks. One of the main highlights is the more expressive regression output, which is easier to understand with some accuracy. In addition, the flexibility of the proposed approach can be enhanced by variants at a very appropriate stage of regression modeling. Experiments are still needed to be able to properly evaluate and optimize correct measurements of both accuracy and resource efficiency.