MileStone1

On

"Store Sales Prediction"

Submitted To

School Of Data Science And Forecasting DAVV, Indore(M.P)

Submitted By :

- Himanshu Jain
- Tarun Choudhary
- Tushar Sonp

Under The Guidance Of :

Prof. Vandit Hedau

Tech Event Organizer :

- Krati Vyas
- Hitesh Kumawat

Date: 02 March 2022

Firstly we take Store Sales data from kaggle (BigMart Sales Data | Kaggle).

From Kaggle we get two types of file

1) Train dataset which consist
    - Number of variable 12
    - Number of Observations 8523 and
2) Test dataset which consist
    - Number Of variables 12
    - Number of Observations 5682.

We start working on jupyter notebook in which we take train.csv file and had done EDA (Expolatory Data Anaylsis) on the dataset.

```
In [12]: import pandas as pd
         import seaborn as sns
         from pandas_profiling import ProfileReport
```

```
In [11]: df = pd.read_csv("Train.csv")
         df.head()
```

Out[11]:

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location
0	FDA15	9.30	Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	
1	DRC01	5.92	Regular	0.019278	Soft Drinks	48.2692	OUT018	2009	Medium	
2	FDN15	17.50	Low Fat	0.016760	Meat	141.6180	OUT049	1999	Medium	
3	FDX07	19.20	Regular	0.000000	Fruits and Vegetables	182.0950	OUT010	1998	NaN	
4	NCD19	8.93	Low Fat	0.000000	Household	53.8614	OUT013	1987	High	

```
In [16]: df.describe()
```

Out[16]:

	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

```
In [17]: df.isnull().sum()
```

```
Out[17]: Item_Identifier               0
         Item_Weight                1463
         Item_Fat_Content              0
         Item_Visibility               0
         Item_Type                     0
         Item_MRP                      0
         Outlet_Identifier             0
         Outlet_Establishment_Year     0
         Outlet_Size                2410
         Outlet_Location_Type          0
         Outlet_Type                   0
         Item Outlet Sales             0
```

Overview

Overview | Alerts 20 | Reproduction

Dataset statistics

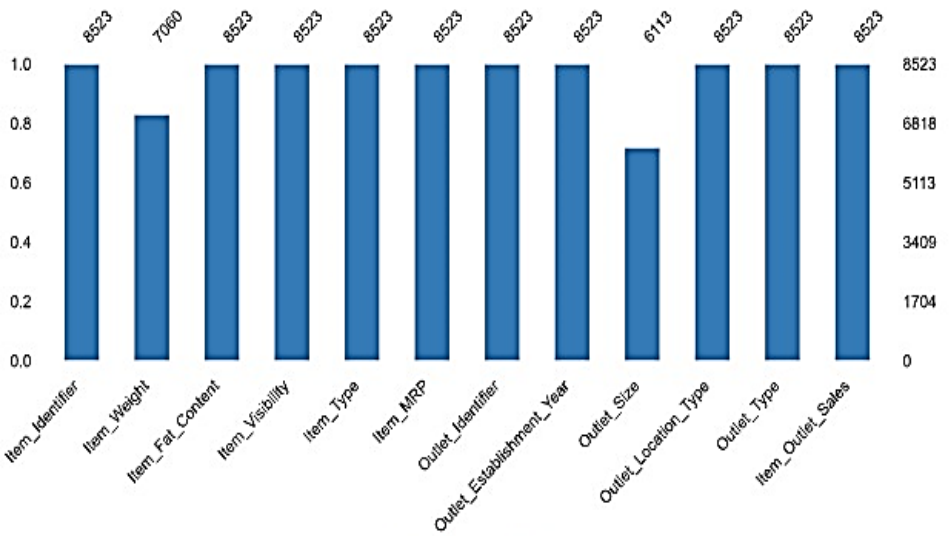
Number of variables	12
Number of observations	8523
Missing cells	3873
Missing cells (%)	3.8%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	799.2 KiB
Average record size in memory	96.0 B

Variable types

Categorical	7
Numeric	5

Missing values

Count | Matrix | Heatmap | Dendrogram



A simple visualization of nullity by column.

Correlations