

Capstone Project Weekly Progress Report

Project Title	Big_Mart Data Visualization and Analysis
Group Name	Group D
Student names/Student IDs	Avik Kundal(744823),Jasmeet Kaur(744215),Kirandeep Kaur(742276),Savreet Kaur(742785),Sukhjinder Singh(743143)
Reporting Week	7 oct 2019 to 13 oct 2019
Faculty Supervisor	William Pourmajidi

1. Tasks Outlined in Previous Weekly Progress Report

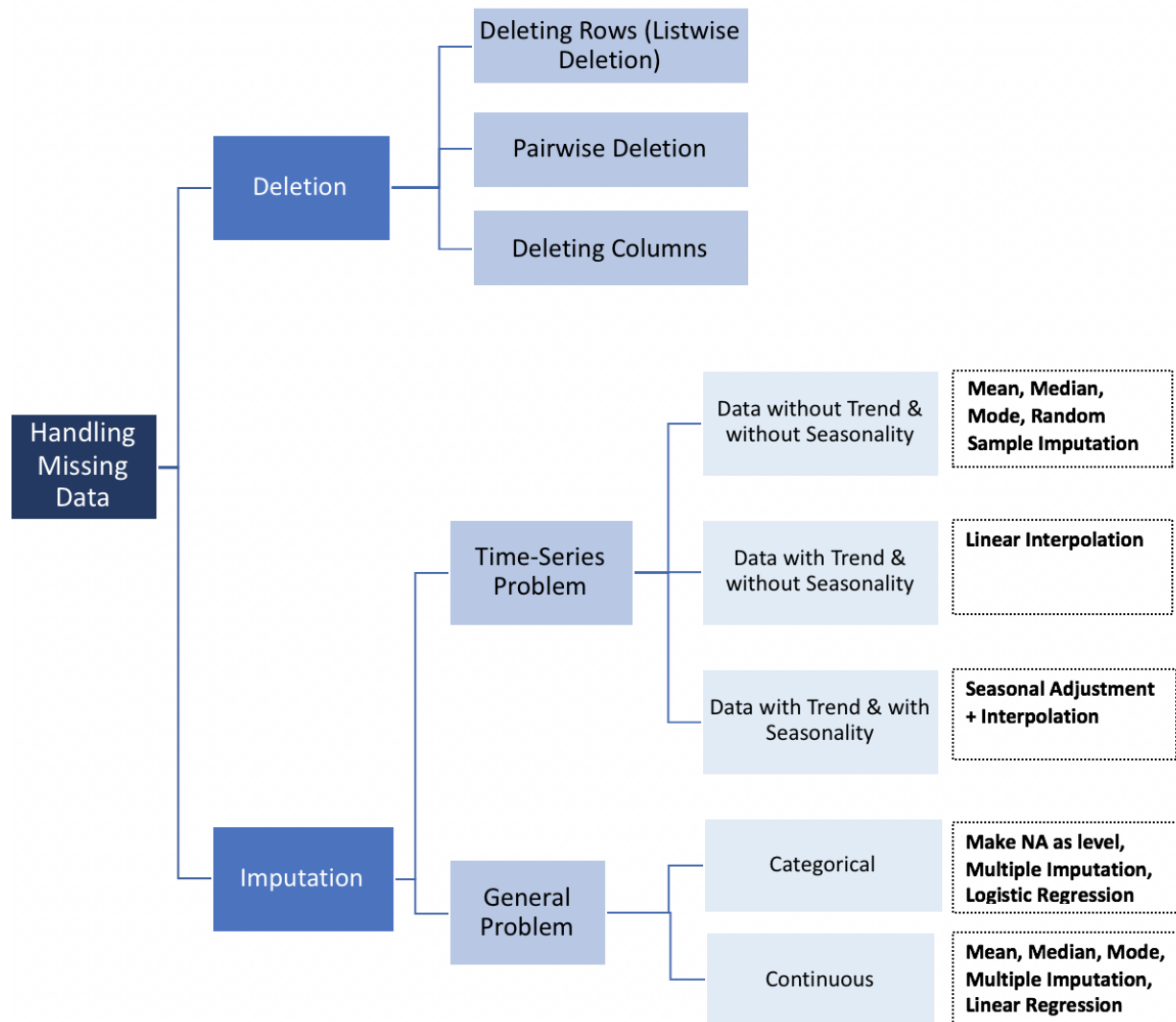
Data Wrangling : Work on missing values or null values present in the data set.

2. Progress Made in Reporting Week

Before jumping to the methods of data imputation, we need to understand the reason why data goes missing.

1. **Missing at Random (MAR):** Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data
2. **Missing Completely at Random (MCAR):** The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.
3. **Missing not at Random (MNAR):** Two possible reasons are that the missing value depends on the hypothetical value (e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable's value (e.g. Let's assume that females generally don't want to reveal their ages! Here the missing value in age variable is impacted by gender variable)

In the first two cases, it is safe to remove the data with missing values depending upon their occurrences, while in the third case removing observations with missing values can produce a bias in the model. So we have to be really careful before removing observations. Note that imputation does not necessarily give better results.



Handling missing values:

- `sum(x.isnull())` – Number of cells with null values in a column
- `apply()` – Apply a function (eg. Apply function on data frame)
- `isnull()` – Get null values present in a column

```

jupyter Big_mart Last Checkpoint: 09/28/2019 (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Python 3

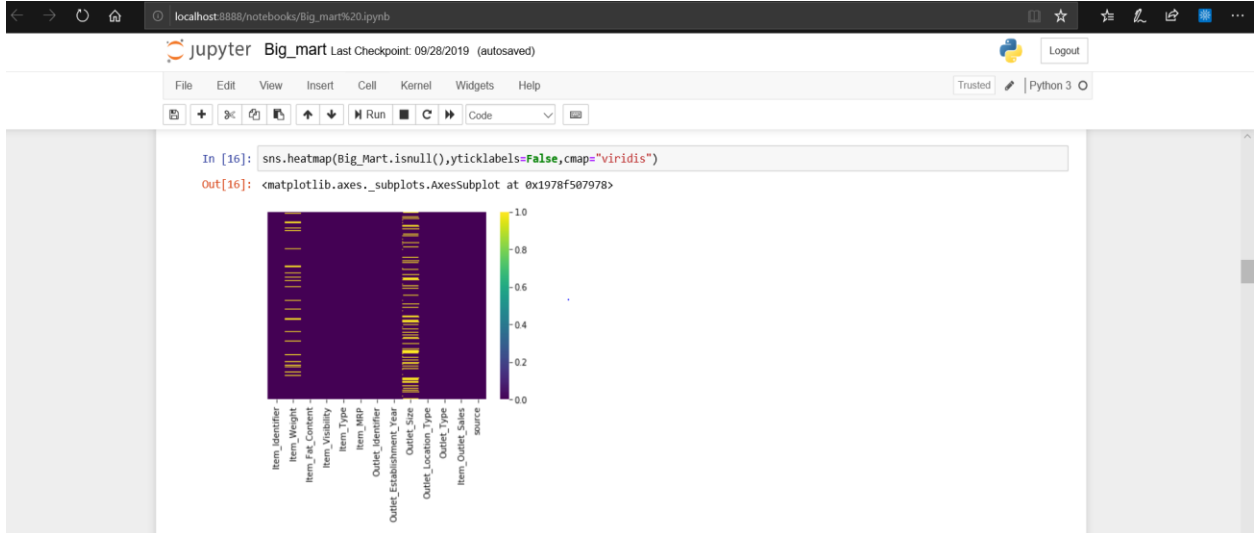
In [14]: Big_Mart.apply(lambda x: sum(x.isnull()))
Out[14]: Item_Identifier      0
         Item_Weight      2439
         Item_Fat_Content    0
         Item_Visibility    0
         Item_Type          0
         Item_MRP          0
         Outlet_Identifier  0
         Outlet_Establishment_Year  0
         Outlet_Size      4016
         Outlet_Location_Type  0
         Outlet_Type        0
         Item_Outlet_Sales    0
         source            0
         dtype: int64

In [15]: Big_Mart.apply(lambda x: len(x.unique()))
Out[15]: Item_Identifier      1559
         Item_Weight        416
         Item_Fat_Content     5
         Item_Visibility    13006
         Item_Type           16
         Item_MRP           8052
         Outlet_Identifier    10
         Outlet_Establishment_Year  9
         Outlet_Size         4
         Outlet_Location_Type  3
         Outlet_Type          4
         Item_Outlet_Sales   3494
         source              2
         dtype: int64

```

Observations: We observed that Item_Weight has 2439 null values Outlet_Size has 4016 null values

The Heatmap below shows that the columns with null values are of different color that is yellow color and rest of heatmap is of purple color with no null values.



```
localhost:8888/notebooks/Big_mart%20.ipynb
jupyter Big_mart Last Checkpoint: 09/28/2019 (autosaved)
Python 3

In [17]: for col in categorical_features:
          print('\n%s column: %s' % (col, Big_Mart[col].value_counts()))

Item_Identifier column:
FDU48      10
FDM48      10
FDM39      10
FDM60      10
FDQ08      10
NCK30      10
FDJ33      10
FDD17      10
FDA50      10
FDL45      10
NCP41      10
FDK02      10
FDT09      10
FDW01      10
FDM40      10
FDT04      10
FDT32      10
...
```

```
localhost:8888/notebooks/Big_mart%20.ipynb
jupyter Big_mart Last Checkpoint: 09/28/2019 (autosaved)
Python 3

In [17]: for col in categorical_features:
          print('\n%s column: %s' % (col, Big_Mart[col].value_counts()))

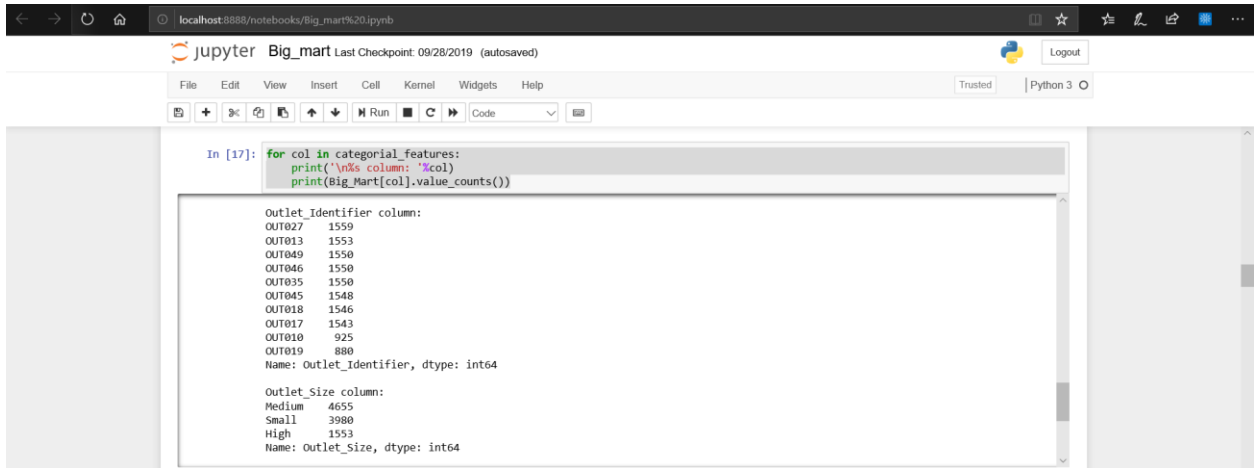
FDM49      7
FDR51      7
FDM10      7
FDM52      7
NCL42      7
DRH11      7
FDM50      7
FDT46      7
FDQ33      7
FDL50      7
Name: Item_Identifier, Length: 1559, dtype: int64

Item_Fat_Content column:
Low Fat      8485
Regular      4824
LF           522
reg          195
low fat      178
Name: Item_Fat_Content, dtype: int64
```

```
localhost:8888/notebooks/Big_mart%20.ipynb
jupyter Big_mart Last Checkpoint: 09/28/2019 (autosaved)
Python 3

In [17]: for col in categorical_features:
          print('\n%s column: %s' % (col, Big_Mart[col].value_counts()))

Item_Type column:
Fruits and Vegetables 2013
Snack Foods           1989
Household             1548
Frozen Foods          1426
Dairy                 1136
Baking Goods          1086
Canned                1084
Health and Hygiene    858
Meat                  736
Soft Drinks           726
Breads                416
Hard Drinks           362
Others                280
Starchy Foods         269
Breakfast             186
Seafood               89
Name: Item_Type, dtype: int64
```



The screenshot shows a Jupyter Notebook titled 'Big_mart' with a last checkpoint from 09/28/2019. The code in cell 17 is as follows:

```
In [17]: for col in categorical_features:
          print('\n%s column: %s' % col)
          print(Big_Mart[col].value_counts())
```

The output of the code is displayed below the cell:

```
Outlet_Identifier column:
OUT027    1559
OUT013    1553
OUT049    1550
OUT046    1550
OUT035    1550
OUT045    1548
OUT018    1546
OUT017    1543
OUT010     925
OUT019     880
Name: Outlet_Identifier, dtype: int64

Outlet_Size column:
Medium    4655
Small     3980
High      1553
Name: Outlet_Size, dtype: int64
```

3. Difficulties Encountered in Reporting Week

To choose proper functions to know about the missing values in Big_Mart data set and to count the categorical features for all the columns present in the data set.

4. Tasks to Be Completed in Next Week

- Replace Null Values with their mean value
- Change the Different names used to represent single item into single name
- For instance, Low Fat and LF in dataset have same meaning. So, replace 'LF' with 'Low Fat' and 'reg' with 'Regular'.