

Capstone Project Weekly Progress Report

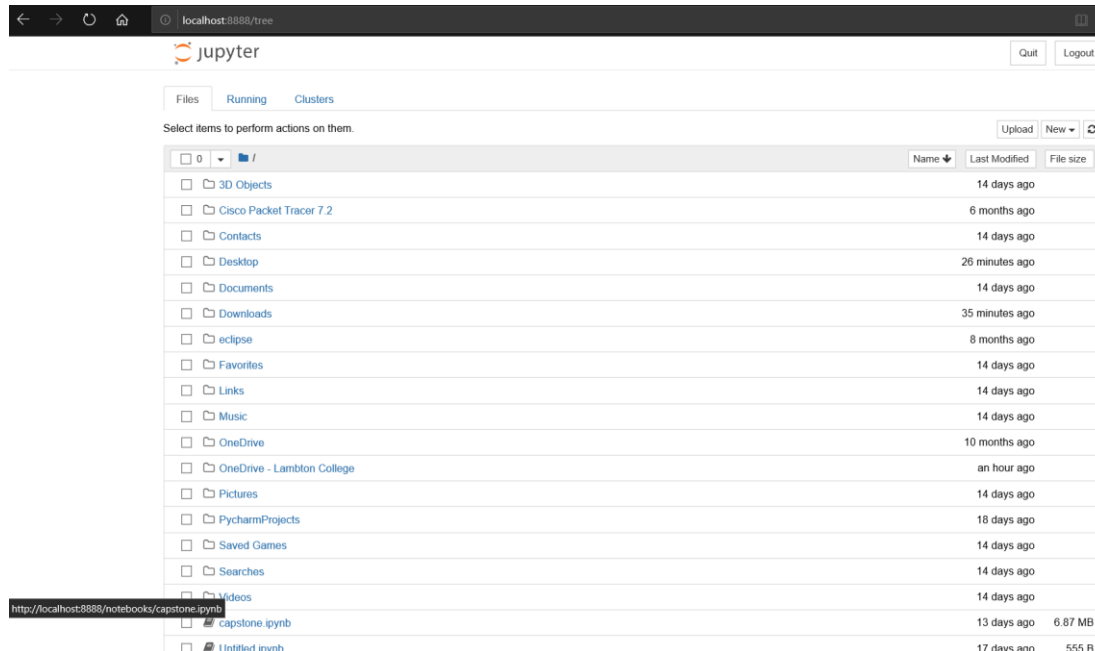
Project Title	Market Based Data Visualization and Analysis
Group Name	Group D
Student names/Student IDs	Avik Kundal(744823),Jasmeet Kaur(744215),Kirandeep Kaur(742276),Savreet Kaur(742785),Sukhjinder Singh(743143)
Reporting Week	23 sept 2019 to 29 sept 2019
Faculty Supervisor	William Pourmajidi

1. Tasks Outlined in Previous Weekly Progress Report (Provide detailed information on the tasks to be completed in this week)

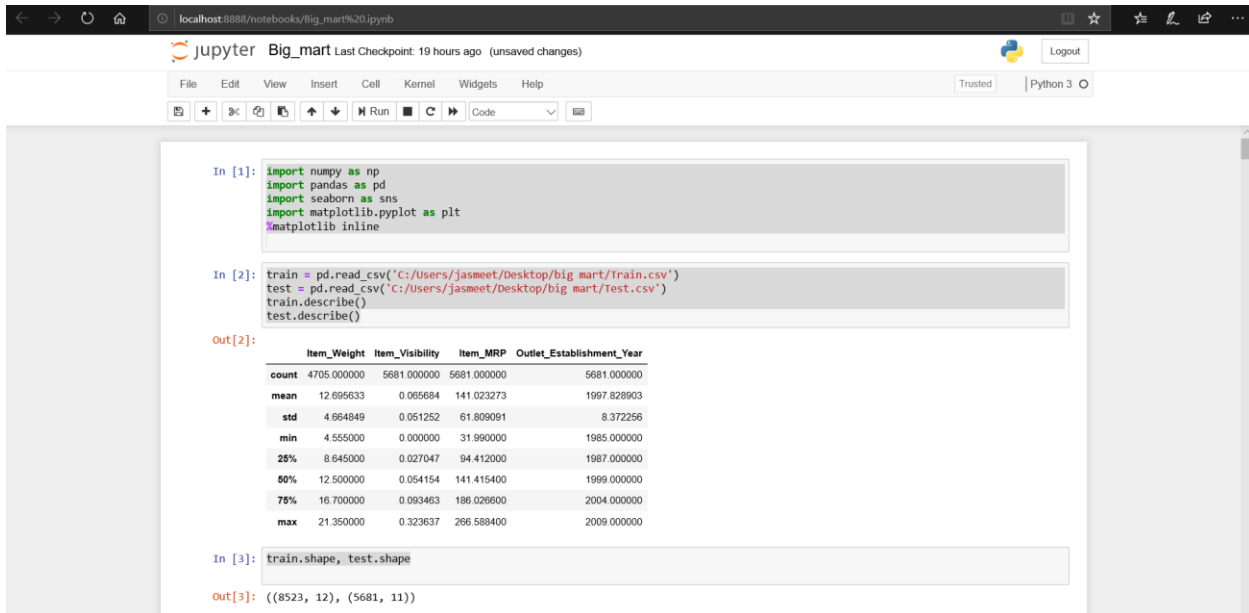
To choose proper libraries of Jupyter Notebook for visualizing Data Set and to concatenate data in test and train sets into a data frame of Pandas library.

2. Progress Made in Reporting Week (Provide detailed information on the progress that you made in the reporting week. Limit your write-up to no more than two page)

- **Home page of Jupyter Notebook**



- **train.describe()** function shows statistics of the training data set.
- **test.describe()** function shows statistics of the test data set.
- **Train.shape, test.shape** shows the number of rows and columns of the data set



```

In [1]: import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
matplotlib inline

In [2]: train = pd.read_csv('C:/Users/jasmeet/Desktop/big mart/Train.csv')
test = pd.read_csv('C:/Users/jasmeet/Desktop/big mart/Test.csv')
train.describe()
test.describe()

Out[2]:

```

	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year
count	4705.000000	5681.000000	5681.000000	5681.000000
mean	12.695633	0.065684	141.023273	1997.828903
std	4.664849	0.051252	61.809091	8.372256
min	4.555000	0.000000	31.990000	1985.000000
25%	8.645000	0.027047	94.412000	1987.000000
50%	12.500000	0.054154	141.415400	1999.000000
75%	16.700000	0.093463	186.026600	2004.000000
max	21.350000	0.323637	266.588400	2009.000000

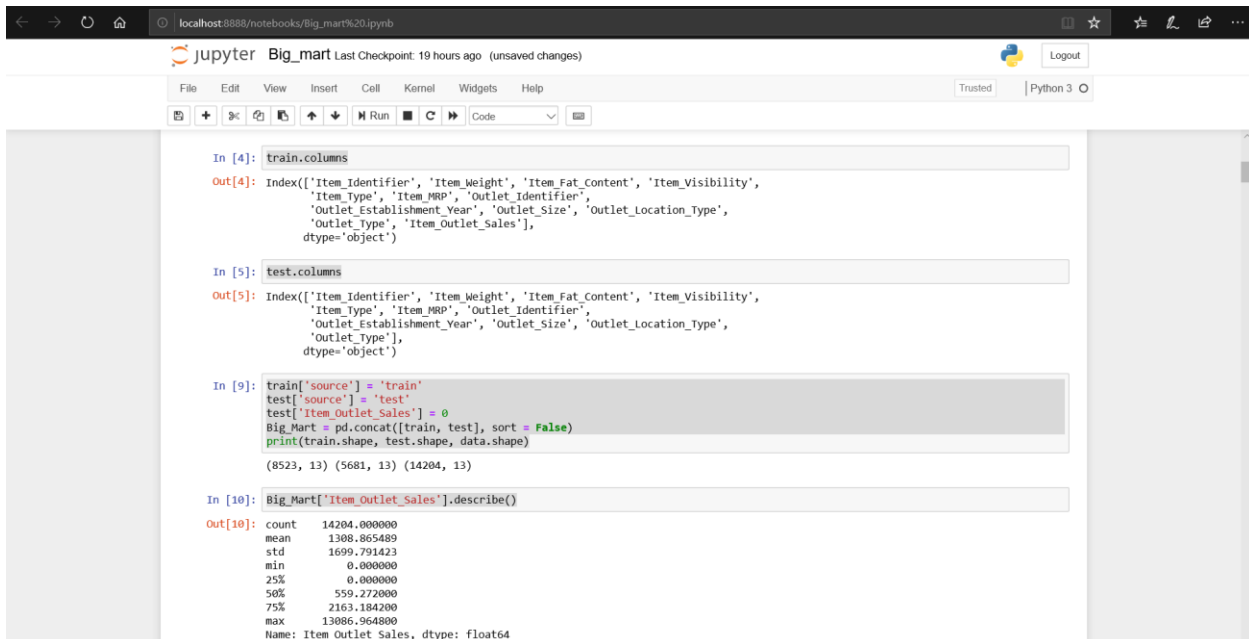
```

In [3]: train.shape, test.shape

Out[3]: ((8523, 12), (5681, 11))

```

- **Train.columns-** Get the columns in train data set
- **Test.columns** – Get the columns in test data set
- **pd.concat([train, test], sort = False)-** to combine train and test data set and get a combined data frame named **Big_Mart**.



```

In [4]: train.columns
Out[4]: Index(['Item_Identifier', 'Item_Weight', 'Item_Fat_Content', 'Item_Visibility',
              'Item_Type', 'Item_MRP', 'Outlet_Identifier',
              'Outlet_Establishment_Year', 'Outlet_Size', 'Outlet_Location_Type',
              'Outlet_Type', 'Item_Outlet_Sales'],
              dtype='object')

In [5]: test.columns
Out[5]: Index(['Item_Identifier', 'Item_Weight', 'Item_Fat_Content', 'Item_Visibility',
              'Item_Type', 'Item_MRP', 'Outlet_Identifier',
              'Outlet_Establishment_Year', 'Outlet_Size', 'Outlet_Location_Type',
              'Outlet_Type'],
              dtype='object')

In [9]: train['source'] = 'train'
test['source'] = 'test'
test['Item_Outlet_Sales'] = 0
Big_Mart = pd.concat([train, test], sort = False)
print(train.shape, test.shape, data.shape)

(8523, 13) (5681, 13) (14204, 13)

In [10]: Big_Mart['Item_Outlet_Sales'].describe()

Out[10]:

```

	Item_Outlet_Sales
count	14204.000000
mean	1308.865489
std	1699.791423
min	0.000000
25%	0.000000
50%	559.272000
75%	2163.184200
max	13086.964800

Name: Item_Outlet_Sales, dtype: float64

3. **Difficulties Encountered in Reporting Week** (Provide detailed information on the difficulties and issues that you encountered in the reporting week. Limit your write-up to no more than one page)

Choosing proper function to count the number of rows and columns in the train and test data set and further to concatenate them into data frame.

4. **Tasks to Be Completed in Next Week** (Outline the tasks to be completed in the following week)

To know the data types present the Big_Mart Data Set and to differentiate the categorical and numerical features.