ONLINE RETAIL PURCHASE

EVANGELINE SHERYL NIVEDHA D

STORE TAB

FEATURES TAB

SALES TAB

Stores tab

➢ This tab contains information about 45 stores
➢ It has store, size and type

Features tab

❑ Store – this attribute represents the unique numbers
❑ Date – this attribute refers to specific week the data is recorded
❑ Fuel price – this attribute represents the cost of fuel in the region specific date or week
❑ Temperature – this attributes represents average temperature in the region during the week or date
❑ Markdown 1-5 – these are anonymized data related to promotional markdown which are discounts or price reduction on product. They are available after nov 2011. they are not available for all stores at all time. Missing data is mentioned as NA
❑ CPI(consumer price index) – change over the time in the price paid by urban consumer for a market basket of consumer goods and service. It reflects the inflation or changes in cost of living
❑ Unemployment – this attribute represents the unemployment rate for the region in specific week
❑ IsHoliday – this attribute represent whether the week is special holiday or public holiday. When people might shop more.

Sales tab

❑ Store – this attribute represents a unique number

❑ Departments – this attribute the department number, indicating which section or category of products the sales data corresponds to

❑ Date- this attribute represent the specific week or date when the sales occurred.

❑ Weekly sales – this attribute represent the amount of money generated from the sales for particular department in a specific store during the given week.

❑ IsHoliday – this attribute represent whether the week is special holiday or public holiday. When people might shop more.

PROCESS INVOLVED

DATA CLEANING

MERGE THE DATA

SUBSETTING AND GROUPING

VISUALIZATION

METHODS INVOLVED

Type	- Explains the type of data involved
Head	- Gives the first top 5 details rowise
Tail	- Gives the last 5 details rowise
Shape	- Gives the number of rows and columns
Columns	- Gives the column labels and their dtype
Info	- Gives the non null count, data type and label
Values	- Gives the number values present in the table
Describe	- Gives the count, mean, max, min 25%, 50% etc
Index	- Gives the range of Index from start to end with step
Sort	- Gives the sorted values based on the Column mentioned
Fillna	- Gives the NA values as 0 in the mentioned dataset
Merge	- Gives the merged output of the datasets mentioned
Subset	- Gives the subset that is to get the specified column rowise as mentioned
Loc	- Gives the output based on the string mentioned, it retrives the data based on the label
Iloc	- Gives the output based on the index and slicing
Groupby	- Gives the grouping based on the column mentioned
Mean	- Gives the mean of the mentioned columns
Reset_Index	- Gives the index from 0 to the final values mentioned

DIAGNOSIS OF DATA AND CLEANING

Type Features

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
1  import pandas as pd
2  import numpy as np
3  import matplotlib.pyplot as plt
4  features=pd.read_csv("Features data set.csv")
5  sales=pd.read_csv("sales data-set.csv")
6  stores=pd.read_csv("stores data-set.csv")
7  print(type(features))
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.
py
<class 'pandas.core.frame.DataFrame'>
PS D:\Assignments\Python>
```

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
2  import numpy as np
3  import matplotlib.pyplot as plt
4  features=pd.read_csv("Features data set.csv")
5  sales=pd.read_csv("sales data-set.csv")
6  stores=pd.read_csv("stores data-set.csv")
7  #print(type(features))
8  print(features.head(5))
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.
py
<class 'pandas.core.frame.DataFrame'>
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.
py
   Store        Date  Temperature  Fuel_Price  ...  MarkDown5         CPI  Unemployment  IsHoliday
0      1  02-05-2010        42.31       2.572  ...        NaN  211.096358         8.106      False
1      1  02-12-2010        38.51       2.548  ...        NaN  211.242170         8.106       True
2      1  02-19-2010        39.93       2.514  ...        NaN  211.289143         8.106      False
3      1  02-26-2010        46.63       2.561  ...        NaN  211.319643         8.106      False
4      1  03-05-2010        46.50       2.625  ...        NaN  211.350143         8.106      False

[5 rows x 12 columns]
PS D:\Assignments\Python>
```

Head of Features

Head of Sales and Stores

```
Boxplot.py   PandasCaseStudy.py
PandasCaseStudy.py > ...
  4   features=pd.read_csv("Features data set.csv")
  5   sales=pd.read_csv("sales data-set.csv")
  6   stores=pd.read_csv("stores data-set.csv")
  7   #print(type(features))
  8   #print(features.head(5))
  9   print(sales.head(5))
 10   print(stores.head(5))

PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.
py
   Store  Dept        Date  Weekly_Sales  IsHoliday
0      1     1  02-05-2010      24924.50      False
1      1     1  02-12-2010      46039.49       True
2      1     1  02-19-2010      41595.55      False
3      1     1  02-26-2010      19403.54      False
4      1     1  03-05-2010      21827.90      False
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.
py
   Store  Dept        Date  Weekly_Sales  IsHoliday
0      1     1  02-05-2010      24924.50      False
1      1     1  02-12-2010      46039.49       True
2      1     1  02-19-2010      41595.55      False
3      1     1  02-26-2010      19403.54      False
4      1     1  03-05-2010      21827.90      False
   Store Type    Size
0      1    A  151315
1      2    A  202307
2      3    B   37392
3      4    A  205863
4      5    B   34875
PS D:\Assignments\Python>
```

Tail of Features

```
Boxplot.py   PandasCaseStudy.py
PandasCaseStudy.py > ...
  5   sales=pd.read_csv("sales data-set.csv")
  6   stores=pd.read_csv("stores data-set.csv")
  7   #print(type(features))
  8   #print(features.head(5))
  9   #print(sales.head(5))
 10   #print(stores.head(5))
 11   print(features.tail(5))

PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

   Store Type    Size
0      1    A  151315
1      2    A  202307
2      3    B   37392
3      4    A  205863
4      5    B   34875
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.
py
      Store        Date  Temperature  Fuel_Price  MarkDown1  ...  MarkDown4  MarkDown5  CPI  Unemployment  IsHoli
day
8185     45  06-28-2013        76.05       3.639     4842.29  ...    2449.97    3169.69  NaN           NaN      Fa
lse
8186     45  07-05-2013        77.50       3.614     9090.48  ...    5797.47    1514.93  NaN           NaN      Fa
lse
8187     45  07-12-2013        79.37       3.614     3789.94  ...     744.84    2150.36  NaN           NaN      Fa
lse
8188     45  07-19-2013        82.84       3.737     2961.49  ...     363.00    1059.46  NaN           NaN      Fa
lse
8189     45  07-26-2013        76.06       3.804      212.02  ...      10.88    1864.57  NaN           NaN      Fa
lse

[5 rows x 12 columns]
PS D:\Assignments\Python>
```

Tail of Stores and Sales

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
 9   #print(sales.head(5))
10   #print(stores.head(5))
11   #print(features.tail(5))
12   print(sales.tail(5))
13   print(stores.tail(5))
14   #features['Date']=pd.to_datetime(features['Date'])
15   #sales['Date']=pd.to_datetime(sales['Date'])
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
[5 rows x 12 columns]
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.
py
        Store  Dept        Date  Weekly_Sales  IsHoliday
421565     45    98  09-28-2012        508.37      False
421566     45    98  10-05-2012        628.10      False
421567     45    98  10-12-2012       1061.02      False
421568     45    98  10-19-2012        760.01      False
421569     45    98  10-26-2012       1076.80      False
    Store Type    Size
40     41    A  196321
41     42    C   39690
42     43    C   41062
43     44    C   39910
44     45    B  118221
PS D:\Assignments\Python>
```

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
13   # print(stores.tail(5))
14   features['Date']=pd.to_datetime(features['Date'])
15   sales['Date']=pd.to_datetime(sales['Date'])
16   print(features['Date'])
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
44     45    B  118221
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy
py
0      2010-02-05
1      2010-02-12
2      2010-02-19
3      2010-02-26
4      2010-03-05
          ...
8185   2013-06-28
8186   2013-07-05
8187   2013-07-12
8188   2013-07-19
8189   2013-07-26
Name: Date, Length: 8190, dtype: datetime64[ns]
PS D:\Assignments\Python>
```

To DateTime

Shapes of Features, Sales and Stores

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
17  #TO KNOW THE NUMBER OF ROWS AND COLUMN
18  print(features.shape)
19  print(sales.shape)
20  print(stores.shape)
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
8186   2013-07-05
8187   2013-07-12
8188   2013-07-19
8189   2013-07-26
Name: Date, Length: 8190, dtype: datetime64[ns]
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.
py
(8190, 12)
(421570, 5)
(45, 3)
PS D:\Assignments\Python>
```

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
21  features.info()
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
 1   Type     45 non-null     object
 2   Size     45 non-null     int64
dtypes: int64(2), object(1)
memory usage: 1.2+ KB
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.
py
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8190 entries, 0 to 8189
Data columns (total 12 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Store          8190 non-null   int64
 1   Date           8190 non-null   object
 2   Temperature    8190 non-null   float64
 3   Fuel_Price     8190 non-null   float64
 4   MarkDown1      4032 non-null   float64
 5   MarkDown2      2921 non-null   float64
 6   MarkDown3      3613 non-null   float64
 7   MarkDown4      3464 non-null   float64
 8   MarkDown5      4050 non-null   float64
 9   CPI            7605 non-null   float64
 10  Unemployment   7605 non-null   float64
 11  IsHoliday      8190 non-null   bool
dtypes: bool(1), float64(9), int64(1), object(1)
memory usage: 712.0+ KB
PS D:\Assignments\Python>
```

Info of Features

Columns of Features, Sales and Stores

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
14  # features['Date']=pd.to_datetime(features['Date'])
15  # sales['Date']=pd.to_datetime(sales['Date'])
16  # print(features['Date'])
17  #TO KNOW THE NUMBER OF ROWS AND COLUMN
18  # print(features.shape)
19  # print(sales.shape)
20  # print(stores.shape)
21  #features.info()
22  #sales.info()
23  #stores.info()
24  print("\n Columns in Feature table\n", features.columns)
25  print("\n Columns in Stores table\n", stores.columns)
26  print("\n Columns in Sales table\n", sales.columns)
27  #Merge the Data in a Unique DataFrame
```

PROBLEMS | OUTPUT | DEBUG CONSOLE | TERMINAL | PORTS

```
 Columns in Feature table
 Index(['Store', 'Date', 'Temperature', 'Fuel_Price', 'MarkDown1', 'MarkDown2',
        'MarkDown3', 'MarkDown4', 'MarkDown5', 'CPI', 'Unemployment',
        'IsHoliday'],
       dtype='object')

 Columns in Stores table
 Index(['Store', 'Type', 'Size'], dtype='object')

 Columns in Sales table
 Index(['Store', 'Dept', 'Date', 'Weekly_Sales', 'IsHoliday'], dtype='object')
PS D:\Assignments\Python>
```

MERGE THE DATA IN UNIQUE DATA FRAME

Merge using features and sales left merge

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
27  #Merge the Data in a Unique DataFrame
28  df=pd.merge(sales,features, on =['Store','Date','IsHoliday'], how='left')
29  print(df)
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.py
        Store  Dept        Date  Weekly_Sales  IsHoliday  ...  MarkDown3  MarkDown4  MarkDown5         CPI  Unemployment
0           1     1  02-05-2010      24924.50      False  ...        NaN        NaN        NaN  211.096358         8.106
1           1     1  02-12-2010      46039.49       True  ...        NaN        NaN        NaN  211.242170         8.106
2           1     1  02-19-2010      41595.55      False  ...        NaN        NaN        NaN  211.289143         8.106
3           1     1  02-26-2010      19403.54      False  ...        NaN        NaN        NaN  211.319643         8.106
4           1     1  03-05-2010      21827.90      False  ...        NaN        NaN        NaN  211.350143         8.106
...       ...   ...         ...           ...        ...  ...        ...        ...        ...         ...           ...
421565     45    98  09-28-2012        508.37      False  ...       1.50    1601.01    3288.25  192.013558         8.684
421566     45    98  10-05-2012        628.10      False  ...      18.82    2253.43    2340.01  192.170412         8.667
421567     45    98  10-12-2012       1061.02      False  ...       7.89     599.32    3990.54  192.327265         8.667
421568     45    98  10-19-2012        760.01      False  ...       3.18     437.73    1537.49  192.330854         8.667
421569     45    98  10-26-2012       1076.80      False  ...     100.00     211.94     858.33  192.308899         8.667

[421570 rows x 14 columns]
PS D:\Assignments\Python>
```

Left Merge using stores, sales and features

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
28  df=pd.merge(sales,features, on =['Store','Date','IsHoliday'], how='left')
29  #print(df)
30  df=pd.merge(df,stores, on=['Store'], how='left')
31  print(df)
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

Python

```
...       ...   ...         ...          ...       ...  ...        ...         ...         ...          ...          ...
421565     45    98  09-28-2012       508.37     False  ...       1.50     1601.01     3288.25  192.013558        8.684
421566     45    98  10-05-2012       628.10     False  ...      18.82     2253.43     2340.01  192.170412        8.667
421567     45    98  10-12-2012      1061.02     False  ...       7.89      599.32     3990.54  192.327265        8.667
421568     45    98  10-19-2012       760.01     False  ...       3.18      437.73     1537.49  192.330854        8.667
421569     45    98  10-26-2012      1076.80     False  ...     100.00      211.94      858.33  192.308899        8.667

[421570 rows x 14 columns]
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.py
        Store  Dept        Date  Weekly_Sales  IsHoliday  ...  MarkDown5         CPI  Unemployment  Type    Size
0           1     1  02-05-2010      24924.50      False  ...        NaN  211.096358         8.106     A  151315
1           1     1  02-12-2010      46039.49       True  ...        NaN  211.242170         8.106     A  151315
2           1     1  02-19-2010      41595.55      False  ...        NaN  211.289143         8.106     A  151315
3           1     1  02-26-2010      19403.54      False  ...        NaN  211.319643         8.106     A  151315
4           1     1  03-05-2010      21827.90      False  ...        NaN  211.350143         8.106     A  151315
...       ...   ...         ...          ...        ...  ...        ...         ...           ...   ...     ...
421565     45    98  09-28-2012       508.37      False  ...    3288.25  192.013558         8.684     B  118221
421566     45    98  10-05-2012       628.10      False  ...    2340.01  192.170412         8.667     B  118221
421567     45    98  10-12-2012      1061.02      False  ...    3990.54  192.327265         8.667     B  118221
421568     45    98  10-19-2012       760.01      False  ...    1537.49  192.330854         8.667     B  118221
421569     45    98  10-26-2012      1076.80      False  ...     858.33  192.308899         8.667     B  118221

[421570 rows x 16 columns]
PS D:\Assignments\Python>
```

Sort and filling of NA of Sales, Stores and Features

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
32  df=df.fillna(0)
33  m=df.sort_values(['Date',"Weekly_Sales"], ascending=[True,False])
34  print(m)
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS — Python

```
3          1    1  02-26-2010      19403.54      False  ...        NaN  211.319643         8.106     A  151315
4          1    1  03-05-2010      21827.90      False  ...        NaN  211.350143         8.106     A  151315
...      ...  ...         ...           ...        ...  ...  ...      ...         ...           ...   ...     ...
421565    45   98  09-28-2012        508.37      False  ...    3288.25  192.013558         8.684     B  118221
421566    45   98  10-05-2012        628.10      False  ...    2340.01  192.170412         8.667     B  118221
421567    45   98  10-12-2012       1061.02      False  ...    3990.54  192.327265         8.667     B  118221
421568    45   98  10-19-2012        760.01      False  ...    1537.49  192.330854         8.667     B  118221
421569    45   98  10-26-2012       1076.80      False  ...     858.33  192.308899         8.667     B  118221

[421570 rows x 16 columns]
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.py
        Store  Dept        Date  Weekly_Sales  IsHoliday  ...  MarkDown5         CPI  Unemployment  Type    Size
137306     14    92  01-06-2012     206871.52      False  ...   15911.56  189.194056         8.424     A  200898
38847       4    92  01-06-2012     183928.47      False  ...    8682.95  130.157516         4.607     A  205863
196731     20    92  01-06-2012     179795.84      False  ...    5460.86  212.571112         6.961     A  203742
127136     13    92  01-06-2012     178257.82      False  ...    7481.58  130.157516         6.104     A  219622
19539       2    92  01-06-2012     177356.35      False  ...    7103.97  219.355063         7.057     A  202307
...       ...   ...         ...           ...        ...  ...  ...      ...         ...           ...   ...     ...
143986     15    47  12-31-2010        -89.00       True  ...       0.00  132.815032         8.067     B  123737
395566     42    72  12-31-2010       -239.00       True  ...       0.00  127.087677         9.003     C   39690
402335     43    72  12-31-2010       -342.84       True  ...       0.00  203.417684        10.210     C   41062
183071     19    47  12-31-2010       -449.00       True  ...       0.00  132.815032         8.067     A  203819
309907     32    47  12-31-2010       -698.00       True  ...       0.00  191.255700         9.137     A  203007

[421570 rows x 16 columns]
PS D:\Assignments\Python>
```

Head, Shape and Columns of the df

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
32  df=df.fillna(0)
33  m=df.sort_values(['Date',"Weekly_Sales"], ascending=[True,False])
34  #print(m)
35  print(df.head(7))
36  print(df.shape)
37  print(df.columns)
38  # #print(df.describe())
39  # #print(df.index)
40  # #print(df.values)
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS — Python

```
   Store  Dept        Date  Weekly_Sales  IsHoliday  Temperature  ...  MarkDown4  MarkDown5         CPI  Unemployment Type    S
ize
0      1     1  02-05-2010      24924.50      False        42.31  ...        0.0        0.0  211.096358         8.106    A  151
315
1      1     1  02-12-2010      46039.49       True        38.51  ...        0.0        0.0  211.242170         8.106    A  151
315
2      1     1  02-19-2010      41595.55      False        39.93  ...        0.0        0.0  211.289143         8.106    A  151
315
3      1     1  02-26-2010      19403.54      False        46.63  ...        0.0        0.0  211.319643         8.106    A  151
315
4      1     1  03-05-2010      21827.90      False        46.50  ...        0.0        0.0  211.350143         8.106    A  151
315
5      1     1  03-12-2010      21043.39      False        57.79  ...        0.0        0.0  211.380643         8.106    A  151
315
6      1     1  03-19-2010      22136.64      False        54.58  ...        0.0        0.0  211.215635         8.106    A  151
315

[7 rows x 16 columns]
(421570, 16)
Index(['Store', 'Dept', 'Date', 'Weekly_Sales', 'IsHoliday', 'Temperature',
       'Fuel_Price', 'MarkDown1', 'MarkDown2', 'MarkDown3', 'MarkDown4',
       'MarkDown5', 'CPI', 'Unemployment', 'Type', 'Size'],
```

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
34  #print(m)
35  # print(df.head(7))
36  # print(df.shape)
37  # print(df.columns)
38  print(df.describe())
39  print(df.index)
40  print(df.values)
41  # #Subsetting
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS — Python

```
28
min         1.000000       1.000000   -4988.940000      -2.060000  ...       0.000000     126.064000       3.879000   34875.0000
00
25%        11.000000      18.000000    2079.650000      46.680000  ...       0.000000     132.022667       6.891000   93638.0000
00
50%        22.000000      37.000000    7612.030000      62.090000  ...       0.000000     182.318780       7.866000  140167.0000
00
75%        33.000000      74.000000   20205.852500      74.280000  ...    2168.040000     212.416993       8.572000  202505.0000
00
max        45.000000      99.000000  693099.360000     100.140000  ...  108519.280000     227.232807      14.313000  219622.0000
00

[8 rows x 13 columns]
RangeIndex(start=0, stop=421570, step=1)
[[1 1 '02-05-2010' ... 8.106 'A' 151315]
 [1 1 '02-12-2010' ... 8.106 'A' 151315]
 [1 1 '02-19-2010' ... 8.106 'A' 151315]
 ...
 [45 98 '10-12-2012' ... 8.667 'B' 118221]
 [45 98 '10-19-2012' ... 8.667 'B' 118221]
 [45 98 '10-26-2012' ... 8.667 'B' 118221]]
PS D:\Assignments\Python>
```

Describe, Index and Values of df

SUBSETTING

Subset of Sales

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
39  # print(df.index)
40  # print(df.values)
41  #Subsetting
42  subset=sales[['Weekly_Sales','Date']]
43  print(subset.head())
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.
py
   Weekly_Sales        Date
0      24924.50  02-05-2010
1      46039.49  02-12-2010
2      41595.55  02-19-2010
3      19403.54  02-26-2010
4      21827.90  03-05-2010
PS D:\Assignments\Python>
```

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
44  # #print(df.head())
45  print(df.loc[0])
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
                    > & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.
py
Store                 1
Dept                  1
Date         02-05-2010
Weekly_Sales      24924.5
IsHoliday           False
Temperature         42.31
Fuel_Price          2.572
MarkDown1             0.0
MarkDown2             0.0
MarkDown3             0.0
MarkDown4             0.0
MarkDown5             0.0
CPI            211.096358
Unemployment        8.106
Type                    A
Size               151315
Name: 0, dtype: object
PS D:\Assignments\Python>
```

Loc for df

Loc for df as a subset of 0th and 99th row

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
42    # subset=sales[['Weekly_Sales','Date']]
43    # print(subset.head())
44    # #print(df.head())
45    #print(df.loc[0])
46    print(df.loc[[0,99]])
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
Size                151315
Name: 0, dtype: object
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.
py
    Store  Dept        Date  Weekly_Sales  IsHoliday  ...  MarkDown5         CPI  Unemployment  Type    Size
0       1     1  02-05-2010      24924.50      False  ...       0.00  211.096358         8.106     A  151315
99      1     1  12-30-2011      23350.88       True  ...    4735.78  219.535990         7.866     A  151315

[2 rows x 16 columns]
PS D:\Assignments\Python>
```

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
40    # print(df.values)
41    #Subsetting
42    # subset=sales[['Weekly_Sales','Date']]
43    # print(subset.head())
44    # #print(df.head())
45    #print(df.loc[0])
46    #print(df.loc[[0,99]])
47    print(df.iloc[0])
48    #print(df.iloc[-1])
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
Name: 421569, dtype: object
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.
py
Store                 1
Dept                  1
Date         02-05-2010
Weekly_Sales      24924.5
IsHoliday           False
Temperature         42.31
Fuel_Price          2.572
MarkDown1             0.0
MarkDown2             0.0
MarkDown3             0.0
MarkDown4             0.0
MarkDown5             0.0
CPI            211.096358
Unemployment        8.106
Type                    A
Size               151315
Name: 0, dtype: object
PS D:\Assignments\Python>
```

iloc for df based on 0th row values

Loc using string of df

```
Boxplot.py   PandasCaseStudy.py
PandasCaseStudy.py > ...
 49    # df.ix[0]
 50    subset=df.loc[:, ['Store','Date','Weekly_Sales']]
 51    print(subset.head())

PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

MarkDown2              0.0
MarkDown3              0.0
MarkDown4              0.0
MarkDown5              0.0
CPI               211.096358
Unemployment           8.106
Type                       A
Size                  151315
Name: 0, dtype: object
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.
py
   Store        Date  Weekly_Sales
0      1  02-05-2010      24924.50
1      1  02-12-2010      46039.49
2      1  02-19-2010      41595.55
3      1  02-26-2010      19403.54
4      1  03-05-2010      21827.90
PS D:\Assignments\Python>
```

Iloc based on slicing of column 2 and 4

```
Boxplot.py   PandasCaseStudy.py
PandasCaseStudy.py > ...
 45    #print(df.loc[0])
 46    #print(df.loc[[0,99]])
 47    #print(df.iloc[0])
 48    #print(df.iloc[-1])
 49    # df.ix[0]
 50    #subset=df.loc[:, ['Store','Date','Weekly_Sales']]
 51    #print(subset.head())
 52    subset=df.iloc[:, [2,4]]
 53    print(subset.head())
 54    # subset=df.iloc[-5::2, :]
 55    # subset.head()
 56

PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.
py
   Store        Date  Weekly_Sales
0      1  02-05-2010      24924.50
1      1  02-12-2010      46039.49
2      1  02-19-2010      41595.55
3      1  02-26-2010      19403.54
4      1  03-05-2010      21827.90
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.
py
         Date  IsHoliday
0  02-05-2010      False
1  02-12-2010       True
2  02-19-2010      False
3  02-26-2010      False
4  03-05-2010      False
PS D:\Assignments\Python>
```

GROUPED CALCULATION

Calculating the mean for CPI grouped based on Date

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
51  #print(subset.head())
52  #subset=df.iloc[:, [2,4]]
53  #print(subset.head())
54  #subset=df.iloc[-5::2, :]
55  #subset.head()
56  print(df.groupby(['Date']) ['CPI'].mean().head(5))
57
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.py
Date
01-06-2012    173.817585
01-07-2011    167.987248
01-13-2012    173.757899
01-14-2011    168.188194
01-20-2012    174.084573
Name: CPI, dtype: float64
PS D:\Assignments\Python>
```

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
53  #print(subset.head())
54  #subset=df.iloc[-5::2, :]
55  #subset.head()
56  #print(df.groupby(['Date']) ['CPI'].mean().head(5))
57  print(df.groupby(['Store','Date']) [['Weekly_Sales','Unemployment']].mean())
58
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
                Weekly_Sales  Unemployment
Store Date
1     01-06-2012  21532.915556         7.348
      01-07-2011  20065.726111         7.742
      01-13-2012  20557.762958         7.348
      01-14-2011  19591.745915         7.742
      01-20-2012  19101.285479         7.348
...                        ...           ...
45    12-17-2010  16765.415672         8.724
      12-23-2011  21742.257000         8.523
      12-24-2010  24389.304783         8.724
      12-30-2011  12785.347500         8.523
      12-31-2010  10136.659701         8.724

[6435 rows x 2 columns]
PS D:\Assignments\Python>
```

Groupby based on Store and Date to find the mean Weekly Sales and Unemployment

Groupby with mean, reset index and head

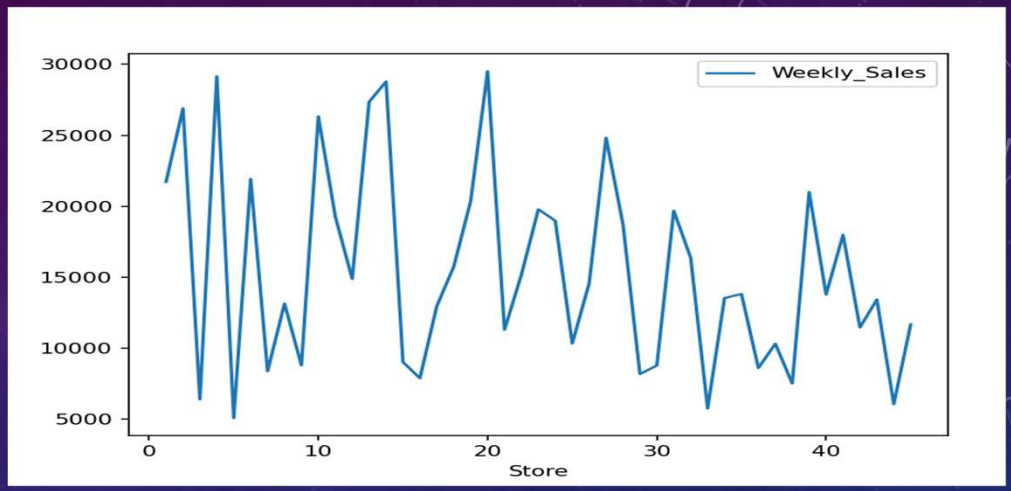
```
Boxplot.py   PandasCaseStudy.py
PandasCaseStudy.py > ...
51   #print(subset.head())
52   #subset=df.iloc[:, [2,4]]
53   #print(subset.head())
54   #subset=df.iloc[-5::2, :]
55   #subset.head()
56   #print(df.groupby(['Date']) ['CPI'].mean().head(5))
57   #print(df.groupby(['Store','Date']) [['Weekly_Sales','Unemployment']].mean())
58   print(df.groupby(['Store','Date']) [['Weekly_Sales','Unemployment']].mean().reset_index().head(5))
59

PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS                    Python

1     01-06-2012  21532.915556         7.348
      01-07-2011  20065.726111         7.742
      01-13-2012  20557.762958         7.348
      01-14-2011  19591.745915         7.742
      01-20-2012  19101.285479         7.348
...                       ...           ...
45    12-17-2010  16765.415672         8.724
      12-23-2011  21742.257000         8.523
      12-24-2010  24389.304783         8.724
      12-30-2011  12785.347500         8.523
      12-31-2010  10136.659701         8.724

[6435 rows x 2 columns]
PS D:\Assignments\Python> & d:/Assignments/Python/myenv/Scripts/python.exe d:/Assignments/Python/PandasCaseStudy.py
   Store        Date  Weekly_Sales  Unemployment
0      1  01-06-2012  21532.915556         7.348
1      1  01-07-2011  20065.726111         7.742
2      1  01-13-2012  20557.762958         7.348
3      1  01-14-2011  19591.745915         7.742
4      1  01-20-2012  19101.285479         7.348
PS D:\Assignments\Python>
```

VISUALIZATION

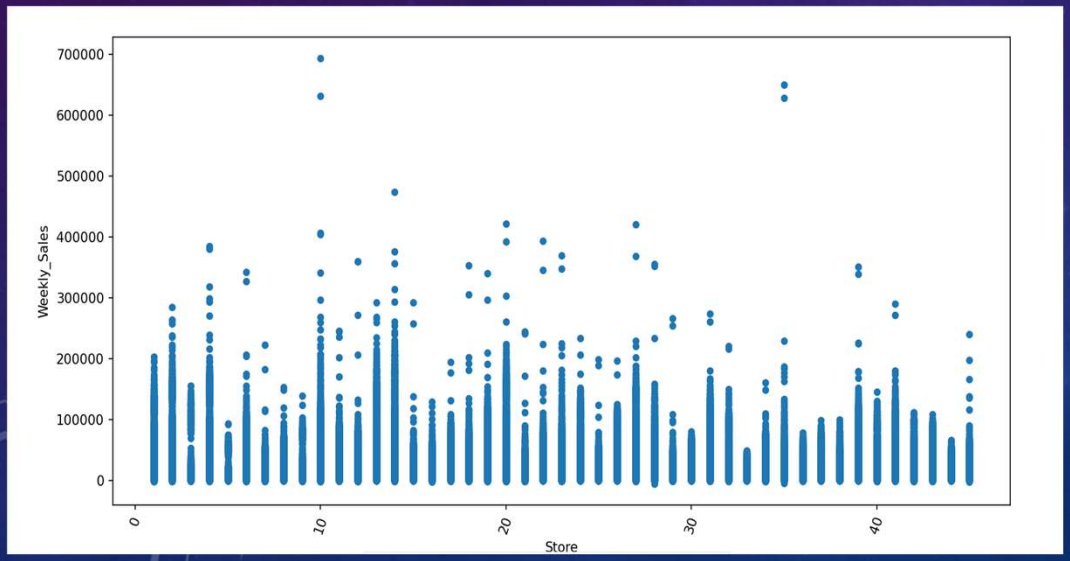
Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
#subset.head()
#print(df.groupby(['Date']) ['CPI'].mean().head(5))
#print(df.groupby(['Store','Date']) [['Weekly_Sales','Unemploymer
#print(df.groupby(['Store','Date']) [['Weekly_Sales','Unemploymer
#VISUALIZATION
(df.groupby(['Store'])[['Weekly_Sales']].mean().plot())
plt.show()
```



Matplotlib using Inline



Matplotlib using Scatter

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

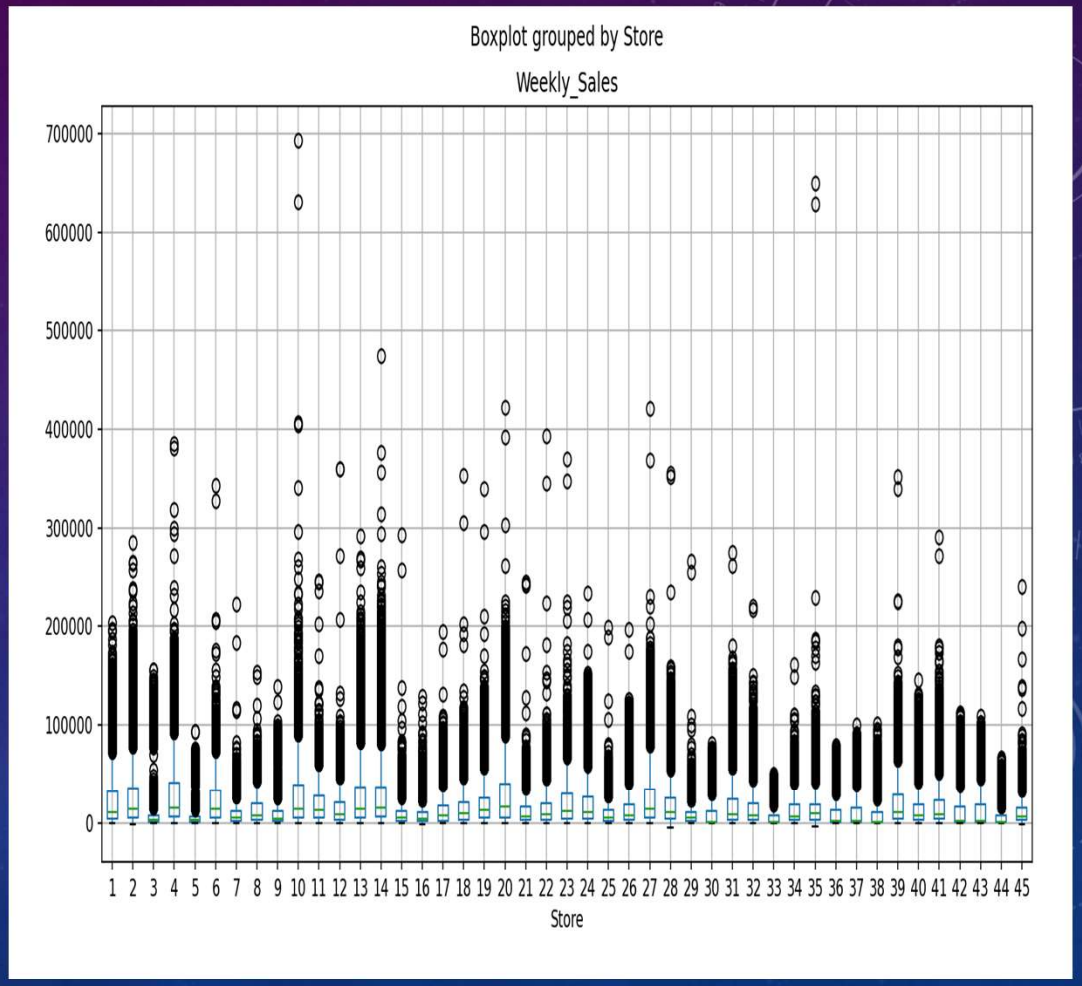
```
#print(df.groupby(['Store','Date']) [['Weekly_Sales','Unemployment'
#VISUALIZATION
#(df.groupby(['Store'])[['Weekly_Sales']].mean().plot())
#plt.show()
df.plot(kind='scatter',x='Store',y='Weekly_Sales',rot=70)
plt.show()
```

Matplotlib using Boxplot

Boxplot.py | PandasCaseStudy.py

PandasCaseStudy.py > ...

```
#VISUALIZATION
#(df.groupby(['Store'])[['Weekly_Sales']].mean().plot())
#plt.show()
#df.plot(kind='scatter',x='Store',y='Weekly_Sales',rot=70)
#plt.show()
df.boxplot(column='Weekly_Sales', by='Store')
plt.show()
```



THANK YOU