EXPERIMENT 3 – DATA CLEANING

AIM: Study basic methods of data cleaning, look at ways of handling missing values, explain data smoothing techniques and discuss approaches to data cleaning as a process.

SOFTWARE REQUIRED:

Spyder IDE 5.1.5

Anaconda3 2021.11 (Python 3.9.7 64-bit)

Anaconda Inc., 2021.11

DATA SET: Real Estate Data Set

PYTHON CODE:

```
import numpy as np # Support for large, multi-dimensional arrays and matrices
import pandas as pd # Library for working with data sets

"""
There is a Unicode character '\u0332', COMBINING LOW LINE*, which acts as an
underline on the character that precedes it in a string. The center() method
will center align the string, using a specified character (space is the
default) as the fill character.
"""

def print_csv_file(df):
    print("\n")
    heading = "Read CSV File"
    print('{:s}'.format('\u0332'.join(heading.center(100))))
    print(df.to_string())

def clean_empty_cells(): # Clean Empty Cells

    def remove_rows(): # Remove Rows

        print("\nRemove Rows - This is usually OK, since data sets can be very
big, and removing a few rows will not have a big impact on the result.")

        change = str(input("Do you want to change the original DataFrame?
(y/n): "))

        if (change == 'N' or change == 'n'): # Return a new Data Frame with no
empty cells
```

```
            new_df = df.dropna() # By default, the dropna() method returns a
new DataFrame, and will not change the original.
            print_csv_file(new_df)

        elif (change == 'Y' or change == 'y'): # Remove all rows with NULL
values
            df.dropna(inplace = True) # Use the inplace = True argument
            print_csv_file(df)
            # Now, the dropna(inplace = True) will NOT return a new DataFrame,
but it will remove all rows containing NULL values from the original
DataFrame.

        else:
            print("Error: Invalid Input! Please try again.")

    def replace_empty_values(): # Replace Empty Values

        print("\nReplace Empty Values - This way you do not have to delete
entire rows just because of some empty cells. The fillna() method allows us to
replace empty cells with a value.")

        value = input("Enter the value with which you want to replace the
empty cell: ")
        df.fillna(value, inplace = True) # Replace NULL values with the number
"value"
        print_csv_file(df)

    def replace_only_specified_columns(): # Replace Only For Specified Columns

        print("\nReplace Only For Specified Columns - To only replace empty
values for one column, specify the column name for the DataFrame.")

        value = input("Enter the value with which you want to replace the
empty cell in the column 'X2. The age of house in years': ")

        df['X2 house age'].fillna(value, inplace = True)
        print_csv_file(df)

    def replace_using_mean(): # Replace Using Mean
        print("\nMean - The average value (the sum of all values divided by
the number of values).")
        mean = df['X2 house age'].mean()
        df['X2 house age'].fillna(mean, inplace = True)
        print_csv_file(df)

    def replace_using_median(): # Replace Using Median
        print("\nMedian - The value in the middle, after you have sorted all
values ascending.")
```

```
        median = df['X2 house age'].median()
        df['X2 house age'].fillna(median, inplace = True)
        print_csv_file(df)

    while True:  # This simulates a Do Loop

        print("\n")
        heading = "CLEAN EMPTY CELLS - MENU"
        print('{:s}'.format('\u0332'.join(heading.center(100))))

        choice = input(
            "   1. Remove Rows\n   2. Replace Empty Values\n   3. Replace Only
For Specified Columns\n   4. Replace Using Mean\n   5. Replace Using
Median\n   6. Exit\nEnter the number corresponding to the menu to implement
the choice: ") # Menu Driven Implementation

        # str() returns the string version of the variable "choice"
        if choice == str(1):
            remove_rows() # Remove Rows
        elif choice == str(2):
            replace_empty_values() # Replace Empty Values
        elif choice == str(3):
            replace_only_specified_columns() # Replace Only For Specified
Columns
        elif choice == str(4):
            replace_using_mean() # Replace Using Mean
        elif choice == str(5):
            replace_using_median() # Replace Using Median
        elif choice == str(6):
            break  # Exit loop
        else:
            print("Error: Invalid Input! Please try again.")

def clean_data_wrong_format(): # Clean Data of Wrong Format

    print("\nIn our Data Frame, we have two cells with the wrong format. Check
out rows 8 and 14, the 'X1 transaction date' column should be a string that
represents a date.")

    def convert_into_correct_format(): # Convert Into a Correct Format

        df['X1 transaction date'] = pd.to_datetime(df['X1 transaction date'])
        print_csv_file(df)

    def remove_rows(): # Remove Rows - The result from the converting in the
example above gave us a NaT value, which can be handled as a NULL value, and
we can remove the row by using the dropna() method.
```

```
        df.dropna(subset=['X1 transaction date'], inplace = True)
        print_csv_file(df)

    while True:  # This simulates a Do Loop

        print("\n")
        heading = "CLEAN DATA OF WRONG FORMAT - MENU"
        print('{:s}'.format('\u0332'.join(heading.center(100))))

        choice = input(
            "   1. Convert Into a Correct Format\n   2. Remove Rows\n   3.
Exit\nEnter the number corresponding to the menu to implement the choice: ") #
Menu Driven Implementation

        # str() returns the string version of the variable "choice".
        if choice == str(1):

            print("\nLet's try to convert all cells in the 'X1. Transaction
Date' column into dates.")
            convert_into_correct_format() # Convert Into a Correct
Format

        elif choice == str(2):

            print("\nAs you can see from the result, the date in row 14 was
fixed, but the empty date in row 8 got a NaT (Not a Time) value, in other
words, an empty value. One way to deal with empty values is simply removing
the entire row.")
            remove_rows() # Remove Rows

        elif choice == str(3):
            break  # Exit loop

        else:
            print("Error: Invalid Input! Please try again.")

def fix_wrong_data(): # Fix Wrong Data

    def convert_into_correct_format(): # Convert Into a Correct Format
        df.loc[4, 'X4 number of convenience stores'] = 10
        print_csv_file(df)

    # Compute the qth quantile of the given data (array elements) along the
specified axis.
    def print_five_number_summary_IQR_outlier(minimum, Q1, median, Q3,
maximum):

        print("Minimum = ", minimum)
```

```
        print("Q1 quantile = ", Q1)
        print("Median =", median)
        print("Q3 quantile = ", Q3)
        print("Maximum =", maximum)

        IQR = Q3 - Q1
        print("Inter-Quartile Range (IQR) = ", IQR)
        outlier = 1.5 * IQR
        print("Outlier (1.5 X IQR) = ", outlier)
        df.loc[4, 'X4 number of convenience stores'] = outlier

    def calc_five_number_summary_variance_standard_deviation():

        min_X4 = df['X4 number of convenience stores'].min()
        Q1_X4 = np.quantile(df['X4 number of convenience stores'], .25)
        median_X4 = df['X4 number of convenience stores'].median()
        Q3_X4 = np.quantile(df['X4 number of convenience stores'], .75)
        max_X4 = df['X4 number of convenience stores'].max()
        print_five_number_summary_IQR_outlier(min_X4, Q1_X4, median_X4, Q3_X4,
max_X4)

        var_X4 = df['X4 number of convenience stores'].var()
        print("Variance = ", var_X4)
        std_X4 = df['X4 number of convenience stores'].std()
        print("Standard Deviation = ", std_X4)

    def remove_rows(): # Remove Rows

        max_value = input("\nEnter the value above which the row should be
deleted: ")
        for i in df.index: # Delete rows where "X4 number of convenience
stores" is higher than "max_value"
            if df.loc[i, 'X4 number of convenience stores'] > int(max_value):
                df.drop(i, inplace = True)
        print_csv_file(df)

    while True:  # This simulates a Do Loop

        print("\n")
        heading = "FIX WRONG - MENU"
        print('{:s}'.format('\u0332'.join(heading.center(100))))

        choice = input(
            "   1. Replace Values\n   2. Remove Rows\n   3. Exit\nEnter the
number corresponding to the menu to implement the choice: ") # Menu Driven
Implementation

        # str() returns the string version of the variable "choice"
```

```
        if choice == str(1):

            print("\nIn our example, it is most likely a typo, and the value
should be '10' instead of '100', and we could just insert '10' in row 5.")
            convert_into_correct_format() # Convert Into a Correct Format

            print("\nFor small data sets, you might be able to replace the
wrong data one by one, but not for big data sets. To replace wrong data for
larger data sets you can create some rules, e.g. outliers.")
            calc_five_number_summary_variance_standard_deviation()
            print_csv_file(df)

        elif choice == str(2):

            print("\nRemove Rows - This way you do not have to find out what
to replace them with, and there is a good chance you do not need them to do
your analyses.")
            remove_rows() # Remove Rows

        elif choice == str(3):
            break  # Exit loop
        else:
            print("Error: Invalid Input! Please try again.")

def remove_duplicates(): # Remove Duplicates

    print("\nTo discover duplicates, we can use the duplicated() method. The
duplicated() method returns a Boolean value for each row:")
    print(df.duplicated()) # Returns True for every row that is a duplicate,
otherwise False.

    print("\nTo remove duplicates, use the drop_duplicates() method.")
    df.drop_duplicates(inplace = True) # Remove all duplicates
    # The (inplace = True) will make sure that the method does NOT return a
new DataFrame, but it will remove all duplicates from the original
DataFrame.
    print_csv_file(df)

# Driver Code: main() ; Execution starts here.

df = pd.read_csv("Real Estate Data Set.csv")
print_csv_file(df)

print("\n")
heading = "Identification of Response Variable & Regressor Variables"
print('{:s}'.format('\u0332'.join(heading.center(100))))

print("\nThere are three regressor variables (X1, X2, X4), namely:")
```

```
print("X1 - Transaction Date")
print("X2 - Age of House in Year(s)")
print("X4 - Number of Convenience Stores within Walking Distance")

print("\n")
heading = "Our Data Set"
print('{:s}'.format('\u0332'.join(heading.center(100))))

print("1. The data set contains some empty cells ('X1 transaction date' in row
9, and 'X2 house age' in rows 7 and 10).")
print("2. The data set contains wrong format ('X1 transaction date' in row
15).")
print("3. The data set contains wrong data ('X4 number of convenience stores'
in row 5).")
print("4. The data set contains duplicates (row 3 and 4).")

while True:  # This simulates a Do Loop

    print("\n")
    heading = "MAIN MENU"
    print('{:s}'.format('\u0332'.join(heading.center(100))))

    choice = input(
        "   1. Clean Empty Cells\n   2. Clean Data of Wrong Format\n   3. Fix
Wrong Data\n   4. Remove Duplicates\n   5. Exit\nEnter the number
corresponding to the menu to implement the choice: ") # Menu Driven
Implementation

    # str() returns the string version of the variable "choice"
    if choice == str(1):
        print("Clean Empty Cells - Empty cells can potentially give you a
wrong result when you analyze data.")
        clean_empty_cells() # Clean Empty Cells

    elif choice == str(2):
        print("Clean Data of Wrong Format - Cells with data of the wrong
format can make it difficult, or even impossible, to analyze data. To fix it,
you have two options:-")
        clean_data_wrong_format() # Clean Data of Wrong Format

    elif choice == str(3):
        fix_wrong_data() # Fix Wrong Data

    elif choice == str(4):
        print("Duplicate rows are rows that have been registered more than one
time.")
        remove_duplicates() # Remove Duplicates
```

```
    elif choice == str(5):
        break # Exit loop

    else:
        print("Error: Invalid Input! Please try again.")
```

RESULT:

Data cleaning routines attempted to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. All the simulation results were verified successfully.

```
Python 3.9.7 (default, Sep 16 2021, 16:59:28) [MSC v.1916 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 7.29.0 -- An enhanced Interactive Python.

Restarting kernel...
```

In [1]:
```
'E:/Plan B/Amrita Vishwa Vidyapeetham/Subject Materials/Semester IV/
19CCE213 - Machine Learning and Artificial Intelligence/Lab/Experiment 3 - Data Cleaning/
Expt_3_Code.py'      ='E:/Plan B/Amrita Vishwa Vidyapeetham/Subject Materials/Semester IV/
19CCE213 - Machine Learning and Artificial Intelligence/Lab/Experiment 3 - Data Cleaning'
```

Read CSV File

	X1 transaction date	X2 house age	X4 number of convenience stores
0	17-09-2012	32.0	10
1	17-09-2012	19.5	9
2	03-08-2013	13.3	5
3	03-08-2013	13.3	5
4	03-03-2012	5.0	100
5	07-06-2012	7.1	3
6	07-06-2012	NaN	7
7	17-04-2013	20.3	6
8	NaN	31.7	1
9	17-04-2013	NaN	3
10	03-08-2013	34.8	1
11	13-03-2013	6.3	9
12	17-09-2012	13.0	5
13	07-06-2012	20.4	4
14	20130506	13.2	4

Identification of Response Variable & Regressor Variables

There are three regressor variables (X1, X2, X4), namely:
X1 - Transaction Date
X2 - Age of House in Year(s)
X4 - Number of Convenience Stores within Walking Distance

Our Data Set

1. The data set contains some empty cells ('X1 transaction date' in row 9, and 'X2 house age' in row 7 and 10).
2. The data set contains wrong format ('X1 transaction date' in row 15).
3. The data set contains wrong data ('X4 number of convenience stores' in row 5).
4. The data set contains duplicates (row 3 and 4).

MAIN MENU

1. Clean Empty Cells
2. Clean Data of Wrong Format
3. Fix Wrong Data

```
   4. Remove Duplicates
   5. Exit
Enter the number corresponding to the menu to implement the choice: 1
Clean Empty Cells - Empty cells can potentially give you a wrong result when you analyze
data.


_____________________________________CLEAN EMPTY CELLS - MENU

   1. Remove Rows
   2. Replace Empty Values
   3. Replace Only For Specified Columns
   4. Replace Using Mean
   5. Replace Using Median
   6. Exit
Enter the number corresponding to the menu to implement the choice: 1

Remove Rows - This is usually OK, since data sets can be very big, and removing a few rows
will not have a big impact on the result.

Do you want to change the original DataFrame? (y/n): y


_____________________________________________Read CSV File
   X1 transaction date  X2 house age  X4 number of convenience stores
0           17-09-2012          32.0                               10
1           17-09-2012          19.5                                9
2           03-08-2013          13.3                                5
3           03-08-2013          13.3                                5
4           03-03-2012           5.0                              100
5           07-06-2012           7.1                                3
7           17-04-2013          20.3                                6
10          03-08-2013          34.8                                1
11          13-03-2013           6.3                                9
12          17-09-2012          13.0                                5
13          07-06-2012          20.4                                4
14            20130506          13.2                                4


_____________________________________CLEAN EMPTY CELLS - MENU

   1. Remove Rows
   2. Replace Empty Values
   3. Replace Only For Specified Columns
   4. Replace Using Mean
   5. Replace Using Median
   6. Exit
Enter the number corresponding to the menu to implement the choice: 7
Error: Invalid Input! Please try again.


_____________________________________CLEAN EMPTY CELLS - MENU

   1. Remove Rows
   2. Replace Empty Values
```

```
   3. Replace Only For Specified Columns
   4. Replace Using Mean
   5. Replace Using Median
   6. Exit
Enter the number corresponding to the menu to implement the choice: 6


_____________________________________________MAIN MENU

   1. Clean Empty Cells
   2. Clean Data of Wrong Format
   3. Fix Wrong Data
   4. Remove Duplicates
   5. Exit
Enter the number corresponding to the menu to implement the choice: 2
Clean Data of Wrong Format - Cells with data of wrong format can make it difficult, or
even impossible, to analyze data. To fix it, you have two options:-

In our Data Frame, we have two cells with the wrong format. Check out row 8 and 14, the
'X1 transaction date' column should be a string that represents a date.


_____________________________CLEAN DATA OF WRONG FORMAT - MENU

   1. Convert Into a Correct Format
   2. Remove Rows
   3. Exit
Enter the number corresponding to the menu to implement the choice: 1

Let's try to convert all cells in the 'X1. Transaction Date' column into dates.


_____________________________________Read CSV File
   X1 transaction date  X2 house age  X4 number of convenience stores
0           2012-09-17          32.0                              10
1           2012-09-17          19.5                               9
2           2013-03-08          13.3                               5
3           2013-03-08          13.3                               5
4           2012-03-03           5.0                             100
5           2012-07-06           7.1                               3
7           2013-04-17          20.3                               6
10          2013-03-08          34.8                               1
11          2013-03-13           6.3                               9
12          2012-09-17          13.0                               5
13          2012-07-06          20.4                               4
14          2013-05-06          13.2                               4


_____________________________CLEAN DATA OF WRONG FORMAT - MENU

   1. Convert Into a Correct Format
   2. Remove Rows
   3. Exit
Enter the number corresponding to the menu to implement the choice: 4
Error: Invalid Input! Please try again.
```

CLEAN DATA OF WRONG FORMAT - MENU

1. Convert Into a Correct Format
2. Remove Rows
3. Exit

Enter the number corresponding to the menu to implement the choice: 3

MAIN MENU

1. Clean Empty Cells
2. Clean Data of Wrong Format
3. Fix Wrong Data
4. Remove Duplicates
5. Exit

Enter the number corresponding to the menu to implement the choice: 3

FIX WRONG - MENU

1. Replace Values
2. Remove Rows
3. Exit

Enter the number corresponding to the menu to implement the choice: 1

In our example, it is most likely a typo, and the value should be '10' instead of '100', and we could just insert '10' in row 5.

Read CSV File

	X1 transaction date	X2 house age	X4 number of convenience stores
0	2012-09-17	32.0	10
1	2012-09-17	19.5	9
2	2013-03-08	13.3	5
3	2013-03-08	13.3	5
4	2012-03-03	5.0	10
5	2012-07-06	7.1	3
7	2013-04-17	20.3	6
10	2013-03-08	34.8	1
11	2013-03-13	6.3	9
12	2012-09-17	13.0	5
13	2012-07-06	20.4	4
14	2013-05-06	13.2	4

For small data sets, you might be able to replace the wrong data one by one, but not for big data sets. To replace wrong data for larger data sets you can create some rules, e.g. outliers.

Minimum = 1
Q1 quantile = 4.0
Median = 5.0
Q3 quantile = 9.0
Maximum = 10
Inter-Quartile Range (IQR) = 5.0

```
Outlier (1.5 X IQR) =  7.5
Variance =  7.2935606060606055
Standard Deviation =  2.700659290999256


_____________________________________________Read CSV File
   X1 transaction date  X2 house age  X4 number of convenience stores
0           2012-09-17          32.0                              10.0
1           2012-09-17          19.5                               9.0
2           2013-03-08          13.3                               5.0
3           2013-03-08          13.3                               5.0
4           2012-03-03           5.0                               7.5
5           2012-07-06           7.1                               3.0
7           2013-04-17          20.3                               6.0
10          2013-03-08          34.8                               1.0
11          2013-03-13           6.3                               9.0
12          2012-09-17          13.0                               5.0
13          2012-07-06          20.4                               4.0
14          2013-05-06          13.2                               4.0


_____________________________________________FIX WRONG - MENU

   1. Replace Values
   2. Remove Rows
   3. Exit
Enter the number corresponding to the menu to implement the choice: 4
Error: Invalid Input! Please try again.


_____________________________________________FIX WRONG - MENU

   1. Replace Values
   2. Remove Rows
   3. Exit
Enter the number corresponding to the menu to implement the choice: 3


_____________________________________________MAIN MENU

   1. Clean Empty Cells
   2. Clean Data of Wrong Format
   3. Fix Wrong Data
   4. Remove Duplicates
   5. Exit
Enter the number corresponding to the menu to implement the choice: 4
Duplicate rows are rows that have been registered more than one time.

To discover duplicates, we can use the duplicated() method. The duplicated() method
returns a Boolean value for each row:
0     False
1     False
2     False
3      True
4     False
```

```
5     False
7     False
10    False
11    False
12    False
13    False
14    False
dtype: bool

To remove duplicates, use the drop_duplicates() method.


                                          Read CSV File
   X1 transaction date  X2 house age  X4 number of convenience stores
0           2012-09-17          32.0                              10.0
1           2012-09-17          19.5                               9.0
2           2013-03-08          13.3                               5.0
4           2012-03-03           5.0                               7.5
5           2012-07-06           7.1                               3.0
7           2013-04-17          20.3                               6.0
10          2013-03-08          34.8                               1.0
11          2013-03-13           6.3                               9.0
12          2012-09-17          13.0                               5.0
13          2012-07-06          20.4                               4.0
14          2013-05-06          13.2                               4.0


                                          MAIN MENU

   1. Clean Empty Cells
   2. Clean Data of Wrong Format
   3. Fix Wrong Data
   4. Remove Duplicates
   5. Exit
Enter the number corresponding to the menu to implement the choice: 6
Error: Invalid Input! Please try again.


                                          MAIN MENU

   1. Clean Empty Cells
   2. Clean Data of Wrong Format
   3. Fix Wrong Data
   4. Remove Duplicates
   5. Exit
Enter the number corresponding to the menu to implement the choice: 5

In [2]:         'E:/Plan B/Amrita Vishwa Vidyapeetham/Subject Materials/Semester IV/
19CCE213 - Machine Learning and Artificial Intelligence/Lab/Experiment 3 - Data Cleaning/
Expt_3_Code.py'      ='E:/Plan B/Amrita Vishwa Vidyapeetham/Subject Materials/Semester IV/
19CCE213 - Machine Learning and Artificial Intelligence/Lab/Experiment 3 - Data Cleaning'


                                          Read CSV File
   X1 transaction date  X2 house age  X4 number of convenience stores
```

```
0           17-09-2012          32.0                          10
1           17-09-2012          19.5                           9
2           03-08-2013          13.3                           5
3           03-08-2013          13.3                           5
4           03-03-2012           5.0                         100
5           07-06-2012           7.1                           3
6           07-06-2012           NaN                           7
7           17-04-2013          20.3                           6
8                  NaN          31.7                           1
9           17-04-2013           NaN                           3
10          03-08-2013          34.8                           1
11          13-03-2013           6.3                           9
12          17-09-2012          13.0                           5
13          07-06-2012          20.4                           4
14            20130506          13.2                           4
```

Identification of Response Variable & Regressor Variables

There are three regressor variables (X1, X2, X4), namely:
X1 - Transaction Date
X2 - Age of House in Year(s)
X4 - Number of Convenience Stores within Walking Distance

Our Data Set

1. The data set contains some empty cells ('X1 transaction date' in row 9, and 'X2 house age' in row 7 and 10).
2. The data set contains wrong format ('X1 transaction date' in row 15).
3. The data set contains wrong data ('X4 number of convenience stores' in row 5).
4. The data set contains duplicates (row 3 and 4).

MAIN MENU

1. Clean Empty Cells
2. Clean Data of Wrong Format
3. Fix Wrong Data
4. Remove Duplicates
5. Exit

Enter the number corresponding to the menu to implement the choice: 1
Clean Empty Cells - Empty cells can potentially give you a wrong result when you analyze data.

CLEAN EMPTY CELLS - MENU

1. Remove Rows
2. Replace Empty Values
3. Replace Only For Specified Columns
4. Replace Using Mean
5. Replace Using Median
6. Exit

Enter the number corresponding to the menu to implement the choice: 1

Remove Rows - This is usually OK, since data sets can be very big, and removing a few rows will not have a big impact on the result.

Do you want to change the original DataFrame? (y/n): n

Read CSV File

	X1 transaction date	X2 house age	X4 number of convenience stores
0	17-09-2012	32.0	10
1	17-09-2012	19.5	9
2	03-08-2013	13.3	5
3	03-08-2013	13.3	5
4	03-03-2012	5.0	100
5	07-06-2012	7.1	3
7	17-04-2013	20.3	6
10	03-08-2013	34.8	1
11	13-03-2013	6.3	9
12	17-09-2012	13.0	5
13	07-06-2012	20.4	4
14	20130506	13.2	4

CLEAN EMPTY CELLS - MENU

1. Remove Rows
2. Replace Empty Values
3. Replace Only For Specified Columns
4. Replace Using Mean
5. Replace Using Median
6. Exit

Enter the number corresponding to the menu to implement the choice: 6

MAIN MENU

1. Clean Empty Cells
2. Clean Data of Wrong Format
3. Fix Wrong Data
4. Remove Duplicates
5. Exit

Enter the number corresponding to the menu to implement the choice: 2
Clean Data of Wrong Format - Cells with data of wrong format can make it difficult, or even impossible, to analyze data. To fix it, you have two options:-

In our Data Frame, we have two cells with the wrong format. Check out row 8 and 14, the 'X1 transaction date' column should be a string that represents a date.

CLEAN DATA OF WRONG FORMAT - MENU

1. Convert Into a Correct Format
2. Remove Rows
3. Exit

Enter the number corresponding to the menu to implement the choice: 2

As you can see from the result, the date in row 14 was fixed, but the empty date in row 8 got a NaT (Not a Time) value, in other words, an empty value. One way to deal with empty values is simply removing the entire row.

```
                                                Read CSV File
   X1 transaction date  X2 house age  X4 number of convenience stores
0           17-09-2012          32.0                                10
1           17-09-2012          19.5                                 9
2           03-08-2013          13.3                                 5
3           03-08-2013          13.3                                 5
4           03-03-2012           5.0                               100
5           07-06-2012           7.1                                 3
6           07-06-2012           NaN                                 7
7           17-04-2013          20.3                                 6
9           17-04-2013           NaN                                 3
10          03-08-2013          34.8                                 1
11          13-03-2013           6.3                                 9
12          17-09-2012          13.0                                 5
13          07-06-2012          20.4                                 4
14            20130506          13.2                                 4
```

```
                                    CLEAN DATA OF WRONG FORMAT - MENU

   1. Convert Into a Correct Format
   2. Remove Rows
   3. Exit
Enter the number corresponding to the menu to implement the choice: 3
```

```
                                                MAIN MENU

   1. Clean Empty Cells
   2. Clean Data of Wrong Format
   3. Fix Wrong Data
   4. Remove Duplicates
   5. Exit
Enter the number corresponding to the menu to implement the choice: 3
```

```
                                            FIX WRONG - MENU

   1. Replace Values
   2. Remove Rows
   3. Exit
Enter the number corresponding to the menu to implement the choice:
```

```
2           03-08-2013          13.3                                5
3           03-08-2013          13.3                                5
4           03-03-2012           5.0                              100
5           07-06-2012           7.1                                3
6           07-06-2012           NaN                                7
7           17-04-2013          20.3                                6
9           17-04-2013           NaN                                3
10          03-08-2013          34.8                                1
11          13-03-2013           6.3                                9
12          17-09-2012          13.0                                5
13          07-06-2012          20.4                                4
14            20130506          13.2                                4
```

CLEAN DATA OF WRONG FORMAT - MENU

```
   1. Convert Into a Correct Format
   2. Remove Rows
   3. Exit
Enter the number corresponding to the menu to implement the choice: 3
```

MAIN MENU

```
   1. Clean Empty Cells
   2. Clean Data of Wrong Format
   3. Fix Wrong Data
   4. Remove Duplicates
   5. Exit
Enter the number corresponding to the menu to implement the choice: 3
```

FIX WRONG - MENU

```
   1. Replace Values
   2. Remove Rows
   3. Exit
Enter the number corresponding to the menu to implement the choice: 2

Remove Rows - This way you do not have to find out what to replace them with, and there is
a good chance you do not need them to do your analyses.


Enter the value above which the row should be deleted: 10
```

Read CSV File

```
   X1 transaction date  X2 house age  X4 number of convenience stores
0           17-09-2012          32.0                               10
1           17-09-2012          19.5                                9
2           03-08-2013          13.3                                5
3           03-08-2013          13.3                                5
5           07-06-2012           7.1                                3
6           07-06-2012           NaN                                7
7           17-04-2013          20.3                                6
```

```
9           17-04-2013          NaN                                3
10          03-08-2013         34.8                                1
11          13-03-2013          6.3                                9
12          17-09-2012         13.0                                5
13          07-06-2012         20.4                                4
14            20130506         13.2                                4
```

FIX WRONG - MENU

```
   1. Replace Values
   2. Remove Rows
   3. Exit
Enter the number corresponding to the menu to implement the choice: 3
```

MAIN MENU

```
   1. Clean Empty Cells
   2. Clean Data of Wrong Format
   3. Fix Wrong Data
   4. Remove Duplicates
   5. Exit
Enter the number corresponding to the menu to implement the choice: 5
```

```
In [3]:          'E:/Plan B/Amrita Vishwa Vidyapeetham/Subject Materials/Semester IV/
19CCE213 - Machine Learning and Artificial Intelligence/Lab/Experiment 3 - Data Cleaning/
Expt_3_Code.py'      ='E:/Plan B/Amrita Vishwa Vidyapeetham/Subject Materials/Semester IV/
19CCE213 - Machine Learning and Artificial Intelligence/Lab/Experiment 3 - Data Cleaning'
```

Read CSV File

	X1 transaction date	X2 house age	X4 number of convenience stores
0	17-09-2012	32.0	10
1	17-09-2012	19.5	9
2	03-08-2013	13.3	5
3	03-08-2013	13.3	5
4	03-03-2012	5.0	100
5	07-06-2012	7.1	3
6	07-06-2012	NaN	7
7	17-04-2013	20.3	6
8	NaN	31.7	1
9	17-04-2013	NaN	3
10	03-08-2013	34.8	1
11	13-03-2013	6.3	9
12	17-09-2012	13.0	5
13	07-06-2012	20.4	4
14	20130506	13.2	4

Identification of Response Variable & Regressor Variables

There are three regressor variables (X1, X2, X4), namely:
X1 - Transaction Date
X2 - Age of House in Year(s)

X4 - Number of Convenience Stores within Walking Distance

Our Data Set

1. The data set contains some empty cells ('X1 transaction date' in row 9, and 'X2 house age' in row 7 and 10).
2. The data set contains wrong format ('X1 transaction date' in row 15).
3. The data set contains wrong data ('X4 number of convenience stores' in row 5).
4. The data set contains duplicates (row 3 and 4).

MAIN MENU

1. Clean Empty Cells
2. Clean Data of Wrong Format
3. Fix Wrong Data
4. Remove Duplicates
5. Exit

Enter the number corresponding to the menu to implement the choice: 1
Clean Empty Cells - Empty cells can potentially give you a wrong result when you analyze data.

CLEAN EMPTY CELLS - MENU

1. Remove Rows
2. Replace Empty Values
3. Replace Only For Specified Columns
4. Replace Using Mean
5. Replace Using Median
6. Exit

Enter the number corresponding to the menu to implement the choice: 1

Remove Rows - This is usually OK, since data sets can be very big, and removing a few rows will not have a big impact on the result.

Do you want to change the original DataFrame? (y/n): s
Error: Invalid Input! Please try again.

CLEAN EMPTY CELLS - MENU

1. Remove Rows
2. Replace Empty Values
3. Replace Only For Specified Columns
4. Replace Using Mean
5. Replace Using Median
6. Exit

Enter the number corresponding to the menu to implement the choice: 2

Replace Empty Values - This way you do not have to delete entire rows just because of some empty cells. The fillna() method allows us to replace empty cells with a value.

Enter the value with which you want to replace the empty cell: 7

```
                                                Read CSV File
   X1 transaction date  X2 house age  X4 number of convenience stores
0           17-09-2012         32.0                               10
1           17-09-2012         19.5                                9
2           03-08-2013         13.3                                5
3           03-08-2013         13.3                                5
4           03-03-2012          5.0                              100
5           07-06-2012          7.1                                3
6           07-06-2012            7                                7
7           17-04-2013         20.3                                6
8                    7         31.7                                1
9           17-04-2013            7                                3
10          03-08-2013         34.8                                1
11          13-03-2013          6.3                                9
12          17-09-2012         13.0                                5
13          07-06-2012         20.4                                4
14            20130506         13.2                                4


                                          CLEAN EMPTY CELLS - MENU

   1. Remove Rows
   2. Replace Empty Values
   3. Replace Only For Specified Columns
   4. Replace Using Mean
   5. Replace Using Median
   6. Exit
Enter the number corresponding to the menu to implement the choice: 6


                                                MAIN MENU

   1. Clean Empty Cells
   2. Clean Data of Wrong Format
   3. Fix Wrong Data
   4. Remove Duplicates
   5. Exit
Enter the number corresponding to the menu to implement the choice: 5

In [4]:         'E:/Plan B/Amrita Vishwa Vidyapeetham/Subject Materials/Semester IV/
19CCE213 - Machine Learning and Artificial Intelligence/Lab/Experiment 3 - Data Cleaning/
Expt_3_Code.py'      ='E:/Plan B/Amrita Vishwa Vidyapeetham/Subject Materials/Semester IV/
19CCE213 - Machine Learning and Artificial Intelligence/Lab/Experiment 3 - Data Cleaning'


                                                Read CSV File
   X1 transaction date  X2 house age  X4 number of convenience stores
0           17-09-2012         32.0                               10
1           17-09-2012         19.5                                9
2           03-08-2013         13.3                                5
3           03-08-2013         13.3                                5
4           03-03-2012          5.0                              100
5           07-06-2012          7.1                                3
6           07-06-2012          NaN                                7
```

```
7           17-04-2013          20.3                            6
8                  NaN          31.7                            1
9           17-04-2013           NaN                            3
10          03-08-2013          34.8                            1
11          13-03-2013           6.3                            9
12          17-09-2012          13.0                            5
13          07-06-2012          20.4                            4
14            20130506          13.2                            4
```

Identification of Response Variable & Regressor Variables

There are three regressor variables (X1, X2, X4), namely:
X1 - Transaction Date
X2 - Age of House in Year(s)
X4 - Number of Convenience Stores within Walking Distance

Our Data Set

1. The data set contains some empty cells ('X1 transaction date' in row 9, and 'X2 house age' in row 7 and 10).
2. The data set contains wrong format ('X1 transaction date' in row 15).
3. The data set contains wrong data ('X4 number of convenience stores' in row 5).
4. The data set contains duplicates (row 3 and 4).

MAIN MENU

1. Clean Empty Cells
2. Clean Data of Wrong Format
3. Fix Wrong Data
4. Remove Duplicates
5. Exit

```
Enter the number corresponding to the menu to implement the choice: 1
```

Clean Empty Cells - Empty cells can potentially give you a wrong result when you analyze data.

CLEAN EMPTY CELLS - MENU

1. Remove Rows
2. Replace Empty Values
3. Replace Only For Specified Columns
4. Replace Using Mean
5. Replace Using Median
6. Exit

```
Enter the number corresponding to the menu to implement the choice: 3
```

Replace Only For Specified Columns - To only replace empty values for one column, specify the column name for the DataFrame.

```
Enter the value with which you want to replace the empty cell in the column 'X2. The age of house in years': 5
```

```
                                        Read CSV File
   X1 transaction date  X2 house age  X4 number of convenience stores
0           17-09-2012         32.0                               10
1           17-09-2012         19.5                                9
2           03-08-2013         13.3                                5
3           03-08-2013         13.3                                5
4           03-03-2012          5.0                              100
5           07-06-2012          7.1                                3
6           07-06-2012            5                                7
7           17-04-2013         20.3                                6
8                  NaN         31.7                                1
9           17-04-2013            5                                3
10          03-08-2013         34.8                                1
11          13-03-2013          6.3                                9
12          17-09-2012         13.0                                5
13          07-06-2012         20.4                                4
14            20130506         13.2                                4
```

```
                                    CLEAN EMPTY CELLS - MENU

   1. Remove Rows
   2. Replace Empty Values
   3. Replace Only For Specified Columns
   4. Replace Using Mean
   5. Replace Using Median
   6. Exit
Enter the number corresponding to the menu to implement the choice: 6
```

```
                                        MAIN MENU

   1. Clean Empty Cells
   2. Clean Data of Wrong Format
   3. Fix Wrong Data
   4. Remove Duplicates
   5. Exit
Enter the number corresponding to the menu to implement the choice: 5
```

```
In [5]:         'E:/Plan B/Amrita Vishwa Vidyapeetham/Subject Materials/Semester IV/
19CCE213 - Machine Learning and Artificial Intelligence/Lab/Experiment 3 - Data Cleaning/
Expt_3_Code.py'      ='E:/Plan B/Amrita Vishwa Vidyapeetham/Subject Materials/Semester IV/
19CCE213 - Machine Learning and Artificial Intelligence/Lab/Experiment 3 - Data Cleaning'
```

```
                                        Read CSV File
   X1 transaction date  X2 house age  X4 number of convenience stores
0           17-09-2012         32.0                               10
1           17-09-2012         19.5                                9
2           03-08-2013         13.3                                5
3           03-08-2013         13.3                                5
4           03-03-2012          5.0                              100
5           07-06-2012          7.1                                3
6           07-06-2012          NaN                                7
7           17-04-2013         20.3                                6
```

```
8                 NaN         31.7                                1
9          17-04-2013          NaN                                3
10         03-08-2013         34.8                                1
11         13-03-2013          6.3                                9
12         17-09-2012         13.0                                5
13         07-06-2012         20.4                                4
14           20130506         13.2                                4
```

Identification of Response Variable & Regressor Variables

There are three regressor variables (X1, X2, X4), namely:
X1 - Transaction Date
X2 - Age of House in Year(s)
X4 - Number of Convenience Stores within Walking Distance

Our Data Set

1. The data set contains some empty cells ('X1 transaction date' in row 9, and 'X2 house age' in row 7 and 10).
2. The data set contains wrong format ('X1 transaction date' in row 15).
3. The data set contains wrong data ('X4 number of convenience stores' in row 5).
4. The data set contains duplicates (row 3 and 4).

MAIN MENU

```
   1. Clean Empty Cells
   2. Clean Data of Wrong Format
   3. Fix Wrong Data
   4. Remove Duplicates
   5. Exit
Enter the number corresponding to the menu to implement the choice: 1
```

Clean Empty Cells - Empty cells can potentially give you a wrong result when you analyze data.

CLEAN EMPTY CELLS - MENU

```
   1. Remove Rows
   2. Replace Empty Values
   3. Replace Only For Specified Columns
   4. Replace Using Mean
   5. Replace Using Median
   6. Exit
Enter the number corresponding to the menu to implement the choice: 4
```

Mean - The average value (the sum of all values divided by number of values).

Read CSV File

```
   X1 transaction date  X2 house age  X4 number of convenience stores
0           17-09-2012     32.000000                                10
1           17-09-2012     19.500000                                 9
2           03-08-2013     13.300000                                 5
```

```
3           03-08-2013     13.300000                                5
4           03-03-2012      5.000000                              100
5           07-06-2012      7.100000                                3
6           07-06-2012     17.684615                                7
7           17-04-2013     20.300000                                6
8                  NaN     31.700000                                1
9           17-04-2013     17.684615                                3
10          03-08-2013     34.800000                                1
11          13-03-2013      6.300000                                9
12          17-09-2012     13.000000                                5
13          07-06-2012     20.400000                                4
14            20130506     13.200000                                4
```

```
_____________________________________CLEAN EMPTY CELLS - MENU

   1. Remove Rows
   2. Replace Empty Values
   3. Replace Only For Specified Columns
   4. Replace Using Mean
   5. Replace Using Median
   6. Exit
Enter the number corresponding to the menu to implement the choice: 6
```

```
_____________________________________________MAIN MENU

   1. Clean Empty Cells
   2. Clean Data of Wrong Format
   3. Fix Wrong Data
   4. Remove Duplicates
   5. Exit
Enter the number corresponding to the menu to implement the choice: 5
```

```
In [6]:          'E:/Plan B/Amrita Vishwa Vidyapeetham/Subject Materials/Semester IV/
19CCE213 - Machine Learning and Artificial Intelligence/Lab/Experiment 3 - Data Cleaning/
Expt_3_Code.py'      ='E:/Plan B/Amrita Vishwa Vidyapeetham/Subject Materials/Semester IV/
19CCE213 - Machine Learning and Artificial Intelligence/Lab/Experiment 3 - Data Cleaning'
```

```
_____________________________________Read CSV File
   X1 transaction date  X2 house age  X4 number of convenience stores
0           17-09-2012          32.0                                10
1           17-09-2012          19.5                                 9
2           03-08-2013          13.3                                 5
3           03-08-2013          13.3                                 5
4           03-03-2012           5.0                               100
5           07-06-2012           7.1                                 3
6           07-06-2012           NaN                                 7
7           17-04-2013          20.3                                 6
8                  NaN          31.7                                 1
9           17-04-2013           NaN                                 3
10          03-08-2013          34.8                                 1
11          13-03-2013           6.3                                 9
12          17-09-2012          13.0                                 5
```

```
13          07-06-2012          20.4                                4
14            20130506          13.2                                4
```

Identification of Response Variable & Regressor Variables

There are three regressor variables (X1, X2, X4), namely:
X1 - Transaction Date
X2 - Age of House in Year(s)
X4 - Number of Convenience Stores within Walking Distance

Our Data Set

1. The data set contains some empty cells ('X1 transaction date' in row 9, and 'X2 house age' in row 7 and 10).
2. The data set contains wrong format ('X1 transaction date' in row 15).
3. The data set contains wrong data ('X4 number of convenience stores' in row 5).
4. The data set contains duplicates (row 3 and 4).

MAIN MENU

1. Clean Empty Cells
2. Clean Data of Wrong Format
3. Fix Wrong Data
4. Remove Duplicates
5. Exit

```
Enter the number corresponding to the menu to implement the choice: 1
Clean Empty Cells - Empty cells can potentially give you a wrong result when you analyze
data.
```

CLEAN EMPTY CELLS - MENU

1. Remove Rows
2. Replace Empty Values
3. Replace Only For Specified Columns
4. Replace Using Mean
5. Replace Using Median
6. Exit

```
Enter the number corresponding to the menu to implement the choice: 5
```

Median - The value in the middle, after you have sorted all values ascending.

Read CSV File

```
   X1 transaction date  X2 house age  X4 number of convenience stores
0           17-09-2012          32.0                                10
1           17-09-2012          19.5                                 9
2           03-08-2013          13.3                                 5
3           03-08-2013          13.3                                 5
4           03-03-2012           5.0                               100
5           07-06-2012           7.1                                 3
6           07-06-2012          13.3                                 7
7           17-04-2013          20.3                                 6
```

```
8                NaN         31.7                                1
9          17-04-2013         13.3                                3
10         03-08-2013         34.8                                1
11         13-03-2013          6.3                                9
12         17-09-2012         13.0                                5
13         07-06-2012         20.4                                4
14           20130506         13.2                                4


 ____________________________________CLEAN EMPTY CELLS - MENU

   1. Remove Rows
   2. Replace Empty Values
   3. Replace Only For Specified Columns
   4. Replace Using Mean
   5. Replace Using Median
   6. Exit
Enter the number corresponding to the menu to implement the choice: 6


 ____________________________________________MAIN MENU

   1. Clean Empty Cells
   2. Clean Data of Wrong Format
   3. Fix Wrong Data
   4. Remove Duplicates
   5. Exit
Enter the number corresponding to the menu to implement the choice: 5

In [7]:
```