EXPERIMENT 2 – MEASURE OF CENTRAL TENDENCY AND DATA DISPERSION

AIM: To introduce typical measures for basic statistical data description and overview data visualization techniques for various kinds of data.

SOFTWARE REQUIRED:

Spyder IDE 5.1.5

Anaconda3 2021.11 (Python 3.9.7 64-bit)

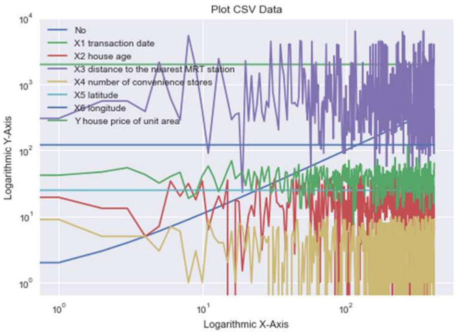
Anaconda Inc., 2021.11

DATA SET: Real Estate Data Set

PYTHON CODE:

```
import matplotlib.pyplot as plt # Provides an implicit way of plotting
import numpy as np # Support for large, multi-dimensional arrays and matrices
import pandas as pd # Library for working with data sets
import seaborn as sb # Provides high-level API to visualize data

# The mean() function returns the values’ mean for the requested axis.
def mean():

    print("\nMean:")

    mean_X2 = df['X2 house age'].mean()
    mean_X3 = df['X3 distance to the nearest MRT station'].mean()
    mean_X4 = df['X4 number of convenience stores'].mean()
    mean_Y = df['Y house price of unit area'].mean()

    print("X2. The age of house in years: " + str(round(mean_X2, 1)))
    print("X3. The distance to nearest MRT station in meters: " +
str(round(mean_X3, 5)))
    print("X4. The number of convenience stores within walking distance: " +
str(round(mean_X4)))
    print("Y. House price per local unit area: " + str(round(mean_Y, 1)))

# The median() method returns a series with the median value of each column.
def median():

    print("\nMedian:")
    median_X2 = df['X2 house age'].median()
    median_X3 = df['X3 distance to the nearest MRT station'].median()
    median_X4 = df['X4 number of convenience stores'].median()
    median_Y = df['Y house price of unit area'].median()

    print("X2. The age of house in years: " + str(median_X2))
    print("X3. The distance to nearest MRT station in meters: " +
str(median_X3))
    print("X4. The number of convenience stores within walking distance: " +
str(int(median_X4)))
    print("Y. House price per local unit area: " + str(median_Y))

# Get each element’s mode(s) along the selected axis.
def mode():

    print("\nMode:")

    mode_X2 = df['X2 house age'].mode()[0]
    mode_X4 = df['X4 number of convenience stores'].mode()[0]

    """
    You can also use mode() to calculate the mode of the sequence, but this
returns a list of numbers, so you'll have to use mode()[0] to get the first
one.
    """

    print("X2. The age of house in years: " + str(int(mode_X2)))
    print("X4. The number of convenience stores within walking distance: " +
str(mode_X4))

    if (int(mode_X2) == 0 or int(mode_X4) == 0):
        print("(Please note that a null value in database is used when the
value in a column is unknown. By default, missing values are not
considered.)")

#  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)

def calc_five_number_summary_variance_standard_deviation():

    print("\nX2. The age of house in years -")

    min_X2 = df['X2 house age'].min()
    Q1_X2 = np.quantile(df['X2 house age'], .25)
    median_X2 = df['X2 house age'].median()
    Q3_X2 = np.quantile(df['X2 house age'], .75)
    max_X2 = df['X2 house age'].max()
    print_five_number_summary_IQR_outlier(min_X2, Q1_X2, median_X2, Q3_X2,
max_X2)

    var_X2 = df['X2 house age'].var()
    print("Variance = ", var_X2)
    std_X2 = df['X2 house age'].std()
    print("Standard Deviation = ", std_X2)

    print("\nX3. The distance to nearest MRT station in meters -")

    min_X3 = df['X3 distance to the nearest MRT station'].min()
    Q1_X3 = np.quantile(df['X3 distance to the nearest MRT station'], .25)
    median_X3 = df['X3 distance to the nearest MRT station'].median()
    Q3_X3 = np.quantile(df['X3 distance to the nearest MRT station'], .75)
    max_X3 = df['X3 distance to the nearest MRT station'].max()
    print_five_number_summary_IQR_outlier(min_X3, Q1_X3, median_X3, Q3_X3,
max_X3)

    var_X3 = df['X3 distance to the nearest MRT station'].var()
    print("Variance = ", var_X3)
    std_X3 = df['X3 distance to the nearest MRT station'].std()
    print("Standard Deviation = ", std_X3)

    print("\nX4. The number of convenience stores within walking distance -")

    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)
    print("\nY. House price per local unit area -")

    min_Y = df['Y house price of unit area'].min()
    Q1_Y = np.quantile(df['Y house price of unit area'], .25)
    median_Y = df['Y house price of unit area'].median()
    Q3_Y = np.quantile(df['Y house price of unit area'], .75)
    max_Y = df['Y house price of unit area'].max()
    print_five_number_summary_IQR_outlier(min_Y, Q1_Y, median_Y, Q3_Y, max_Y)

    var_Y = df['Y house price of unit area'].var()
    print("Variance = ", var_Y)
    std_Y = df['Y house price of unit area'].std()
    print("Standard Deviation = ", std_Y)

def plot_data():

    print("\nPlot CSV Data:")

    """
    The method yscale() or xscale() takes a single value as a parameter which
is the type of conversion of the scale, to convert axes to a logarithmic scale
we pass the “log” keyword or the matplotlib.scale
    """

    plt.xscale('log'); plt.xlabel("Logarithmic X-Axis")
    plt.yscale('log'); plt.ylabel("Logarithmic Y-Axis")
    plt.title("Plot CSV Data")
    plt.plot(df)
    plt.legend(df)
    plt.show()

def boxplot():
    print("\nBoxplot: Graphic display of five-number summary.")
    sb.boxplot(data = df, orient = 'h')
    plt.show()

def histogram():

    print("\nHistogram: X-axis are values and Y-axis represent frequencies.")

    """
    Here, the kde flag is set to False. As a result, the representation of the
kernel estimation plot will be removed and only the histogram is plotted.
    """

    plt.show()
    sb.histplot(df['X2 house age'], kde = False)
    plt.show()
    sb.histplot(df['X3 distance to the nearest MRT station'], kde = False)
    plt.show()
    sb.histplot(df['X4 number of convenience stores'], kde = False)
    plt.show()
    sb.histplot(df['Y house price of unit area'], kde = False)
    plt.show()

def scatter_plot():

    print("\nScatter Plot: Each pair of values is a pair of coordinates and
plotted as points in the plane.")
    house_age = df['X2 house age']
    house_price = df['Y house price of unit area']

    plt.xlabel("X2: The age of house in years")
    plt.ylabel("Y: House price per local unit area")
    plt.title("Relationship Between House Price and House Age")
    plt.scatter(house_age, house_price, cmap="Blues", s=100, alpha=0.6,
edgecolor='black', linewidth=1)

    cbar = plt.colorbar()
    cbar.set_label('Intensity Ratio')
    plt.tight_layout()
    plt.show()

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

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

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

print("\nThere are six regressor variables (from X1 to X6) and one response
variable (namely, y):")
print("No - Serial Number")
print("X1 - Transaction Date")
print("X2 - Age of House in year(s)")
print("X3 - Distance to Nearest MRT station in meter(s)")
print("X4 - Number of Convenience Stores within Walking Distance")
print("X5 - Latitude Coordinates")
print("X6 - Longitude Coordinates")
print("Y - House Price per Local Unit Area")

print("\n")
heading = "Read CSV File"
print('{:s}'.format('\u0332'.join(heading.center(100))))
df = pd.read_csv("Real Estate Data Set.csv"); print(df)

print("\n")
heading = "Measuring the Central Tendency"
print('{:s}'.format('\u0332'.join(heading.center(100))))
mean(); median(); mode()

print("\n")
heading = "Measuring the Dispersion of Data"
print('{:s}'.format('\u0332'.join(heading.center(100))))
calc_five_number_summary_variance_standard_deviation()

print("\n")
heading = "Graphic Displays of Basic Statistical Descriptions"
print('{:s}'.format('\u0332'.join(heading.center(100))))
plt.style.use('seaborn') # To get seaborn type plot
plot_data(); boxplot(); histogram(); scatter_plot()
```

PLOTS:



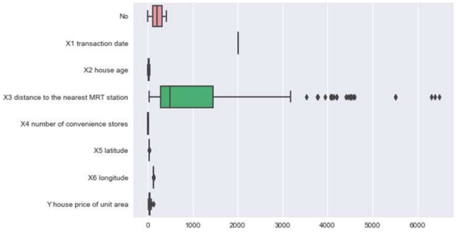
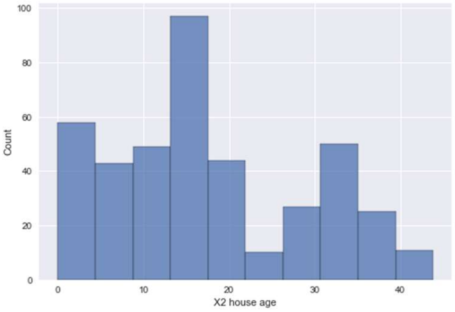


Figure 2. Boxplot



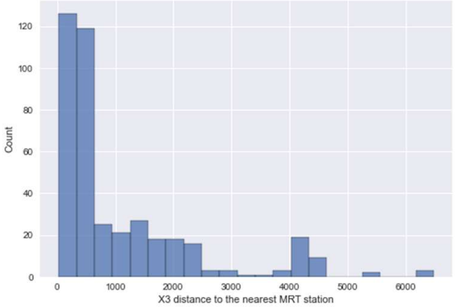


Figure 3. Histogram - X2 house age

Figure 4. Histogram - X3 distance to the nearest MRT station

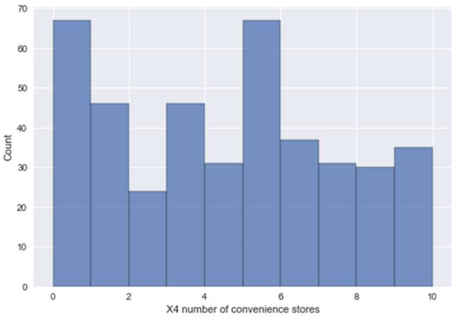


Figure 5. Histogram - X4 number of convenience stores

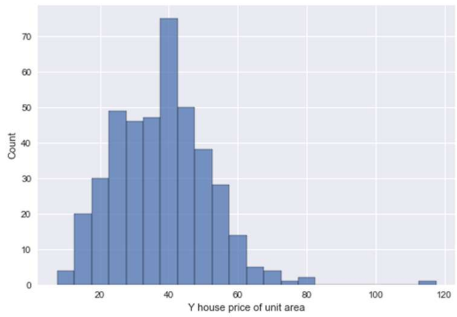


Figure 6. Histogram - Y house price of unit area

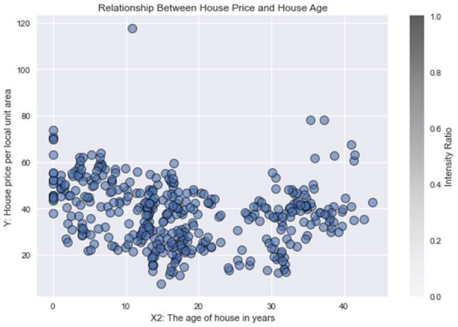


Figure 7. Scatter Plot

RESULT:

Familiarized with typical measures for basic statistical data description and reviewed the data visualization techniques for various kinds of 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 2 - Measure of Central Tendency and Data Dispersion/Expt_2_Code.py'      ='E:/Plan B/Amrita Vishwa Vidyapeetham/Subject Materials/Semester IV/19CCE213 - Machine Learning and Artificial Intelligence/Lab/Experiment 2 - Measure of Central Tendency and Data Dispersion'
```

```
 ____________________Identification of Response Variable & Regressor Variables

There are six regressor variables (from X1 to X6) and one response variable (namely, y):
No - Serial Number
X1 - Transaction Date
X2 - Age of House in year(s)
X3 - Distance to Nearest MRT station in meter(s)
X4 - Number of Convenience Stores within Walking Distance
X5 - Latitude Coordinates
X6 - Longitude Coordinates
Y - House Price per Local Unit Area


 __________________________________________Read CSV File
      No  X1 transaction date  ...  X6 longitude  Y house price of unit area
0      1              2012.917  ...     121.54024                        37.9
1      2              2012.917  ...     121.53951                        42.2
2      3              2013.583  ...     121.54391                        47.3
3      4              2013.500  ...     121.54391                        54.8
4      5              2012.833  ...     121.54245                        43.1
..   ...                   ...  ...           ...                         ...
409  410              2013.000  ...     121.50381                        15.4
410  411              2012.667  ...     121.54310                        50.0
411  412              2013.250  ...     121.53986                        40.6
412  413              2013.000  ...     121.54067                        52.5
413  414              2013.500  ...     121.54310                        63.9

[414 rows x 8 columns]


 __________________________________Measuring the Central Tendency

Mean:
X2. The age of house in years: 17.7
X3. The distance to nearest MRT station in meters: 1083.88569
X4. The number of convenience stores within walking distance: 4
Y. House price per local unit area: 38.0

Median:
```

```
X2. The age of house in years: 16.1
X3. The distance to nearest MRT station in meters: 492.2313
X4. The number of convenience stores within walking distance: 4
Y. House price per local unit area: 38.45

Mode:
X2. The age of house in years: 0
X4. The number of convenience stores within walking distance: 0
(Please note that a null value in database is used when the value in a column is unknown.
By default, missing values are not considered.)
```

Measuring the Dispersion of Data

```
X2. The age of house in years -
Minimum =  0.0
Q1 quantile =  9.025
Median = 16.1
Q3 quantile =  28.15
Maximum = 43.8
Inter-Quartile Range (IQR) =  19.125
Outlier (1.5 X IQR) =  28.6875
Variance =  129.7887038401704
Standard Deviation =  11.392484533242536

X3. The distance to nearest MRT station in meters -
Minimum =  23.38284
Q1 quantile =  289.3248
Median = 492.2313
Q3 quantile =  1454.279
Maximum = 6488.021
Inter-Quartile Range (IQR) =  1164.9542000000001
Outlier (1.5 X IQR) =  1747.4313000000002
Variance =  1592920.6308205703
Standard Deviation =  1262.1095954078514

X4. The number of convenience stores within walking distance -
Minimum =  0
Q1 quantile =  1.0
Median = 4.0
Q3 quantile =  6.0
Maximum = 10
Inter-Quartile Range (IQR) =  5.0
Outlier (1.5 X IQR) =  7.5
Variance =  8.676334350984305
Standard Deviation =  2.945561805663617

Y. House price per local unit area -
Minimum =  7.6
Q1 quantile =  27.7
Median = 38.45
Q3 quantile =  46.6
Maximum = 117.5
Inter-Quartile Range (IQR) =  18.900000000000002
Outlier (1.5 X IQR) =  28.35
```

```
Variance =  185.13650746862245
Standard Deviation =  13.606487697735314
```

Graphic Displays of Basic Statistical Descriptions

Plot CSV Data:

Boxplot: Graphic display of five-number summary.

Histogram: X-axis are values and Y-axis represent frequencies.

Scatter Plot: Each pair of values is a pair of coordinates and plotted as points in the plane.

In [2]: