Introduction to Python Libraries and Dataset Preprocessing

SCENARIO 1: E-Commerce Sales Data

Objective

Analyze product sales trends and detect missing information in an e-commerce dataset.

Steps Performed

- Imported Pandas, NumPy, and Matplotlib
- Loaded dataset into a Pandas DataFrame
- Inspected data using head(), tail(), info(), and describe()
- Checked missing values using isnull().sum()
- Calculated total sales using Quantity × UnitPrice
- Visualized top products using:
    - Bar chart
    - Line chart

Observations

- Missing values were found in the Description and CustomerID columns
- Some products generated significantly higher revenue than others
- Sales distribution was highly skewed with few top-selling products

SCENARIO 2: Hospital Patient Records

Objective

Identify missing health metrics and analyze patient health patterns.

Steps Performed

- Loaded dataset into Pandas
- Explored structure and missing values
- Identified zero values in medical attributes
- Replaced zero values with NaN
- Visualized:
    - Glucose levels using histogram
    - Age distribution using boxplot
- Analyzed mean health metrics grouped by outcome

Observations

- Zero values were present in Glucose, BloodPressure, BMI, Insulin
- Patients with diabetes showed:
    - Higher glucose levels
    - Higher BMI on average
- Age distribution showed wider spread among diabetic patients

SCENARIO 3: Housing Dataset

Objective

Examine missing housing features and relationships affecting house prices.

Steps Performed

- Loaded housing dataset
- Inspected column types and missing values
- Visualized relationships using:
    - Scatter plots
    - Correlation heatmaps

Observations

- Missing values observed in features such as lot size and bedrooms
- Strong correlation found between:
    - House size and price
    - Number of rooms and price
- Visualization helped clearly identify predictors for price estimation

SCENARIO 4: Banking Customer Data

Objective

Understand customer demographics and detect missing banking information.

Steps Performed

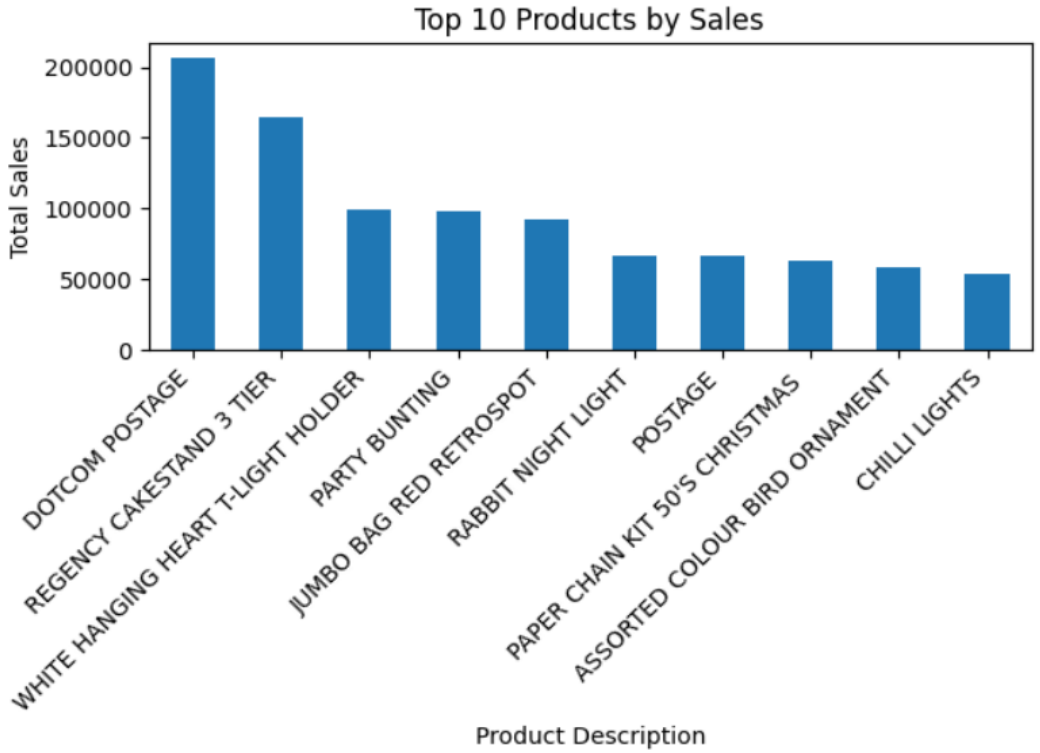
- Imported dataset into Pandas
- Examined data structure and null values
- Visualized:
    - Age distribution (bar plot)
    - Income distribution (box plot)
    - Spending behavior

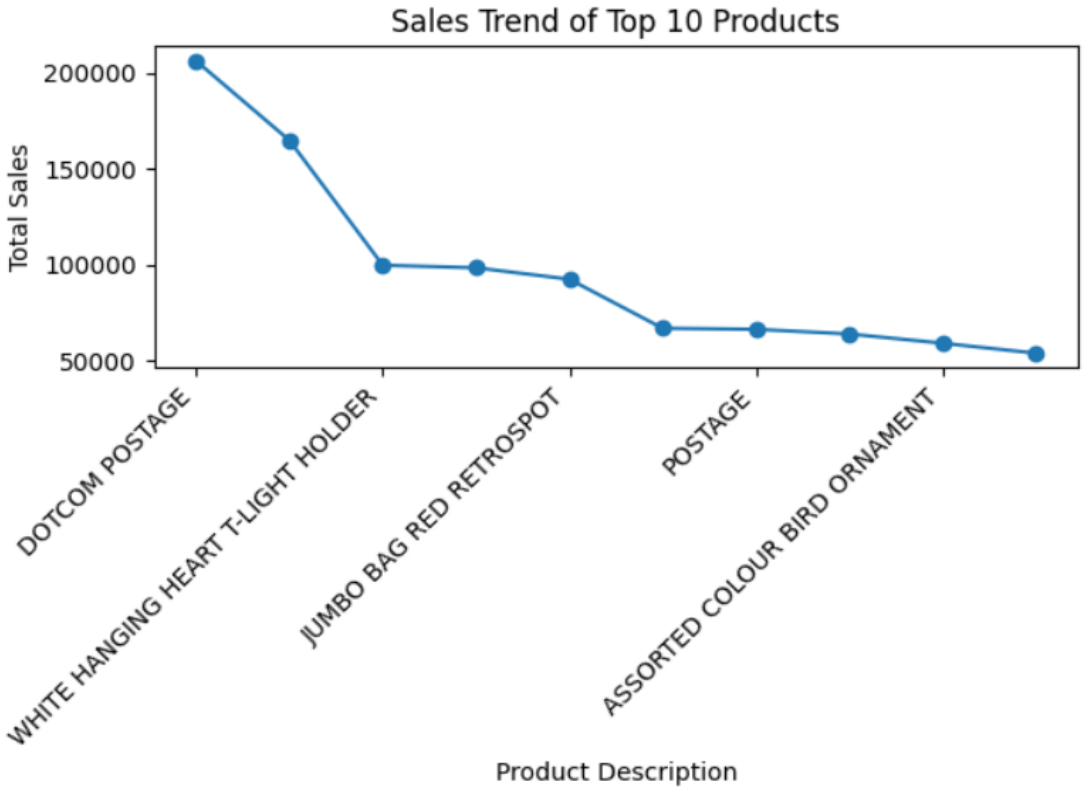
Observations

- Missing values found in Income and Customer profile fields
- Middle-aged customers showed higher spending
- Higher income customers tended to spend more on premium products

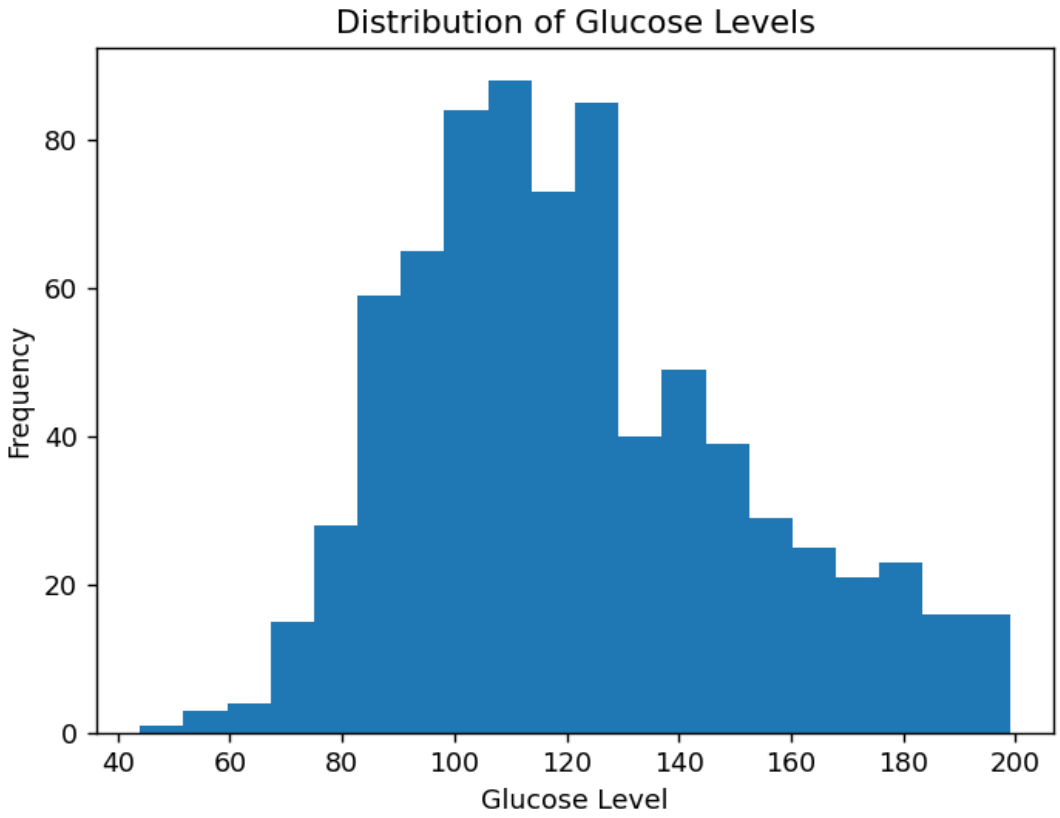
OUTPUT :

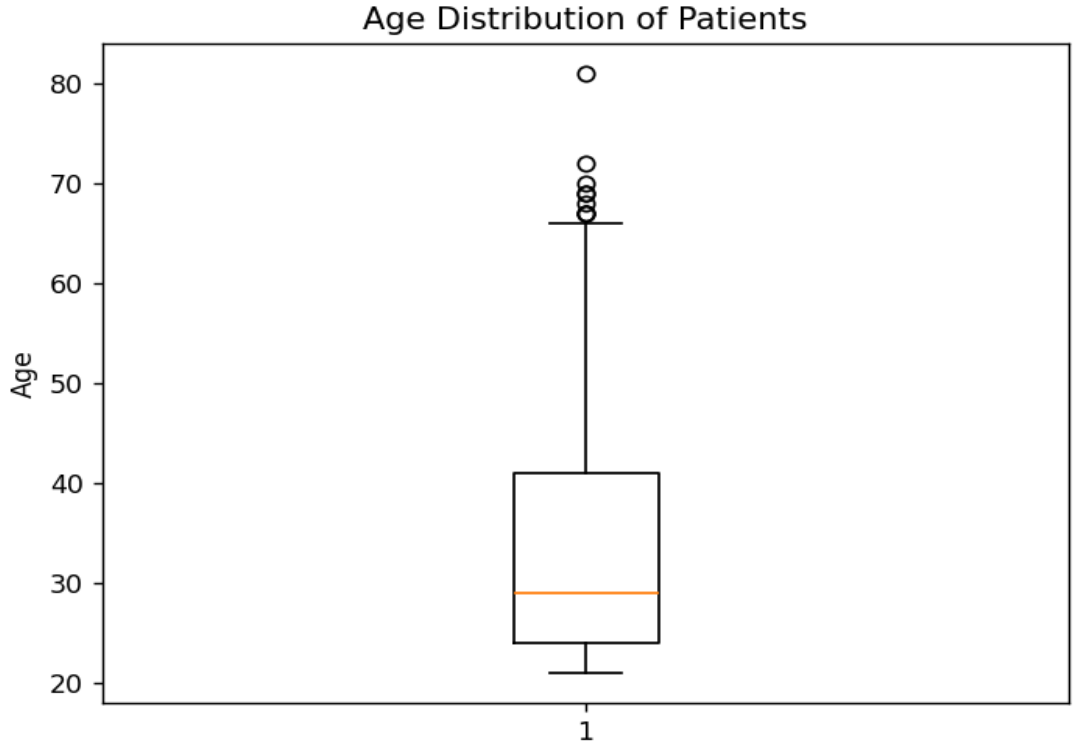
Scenario – 1:



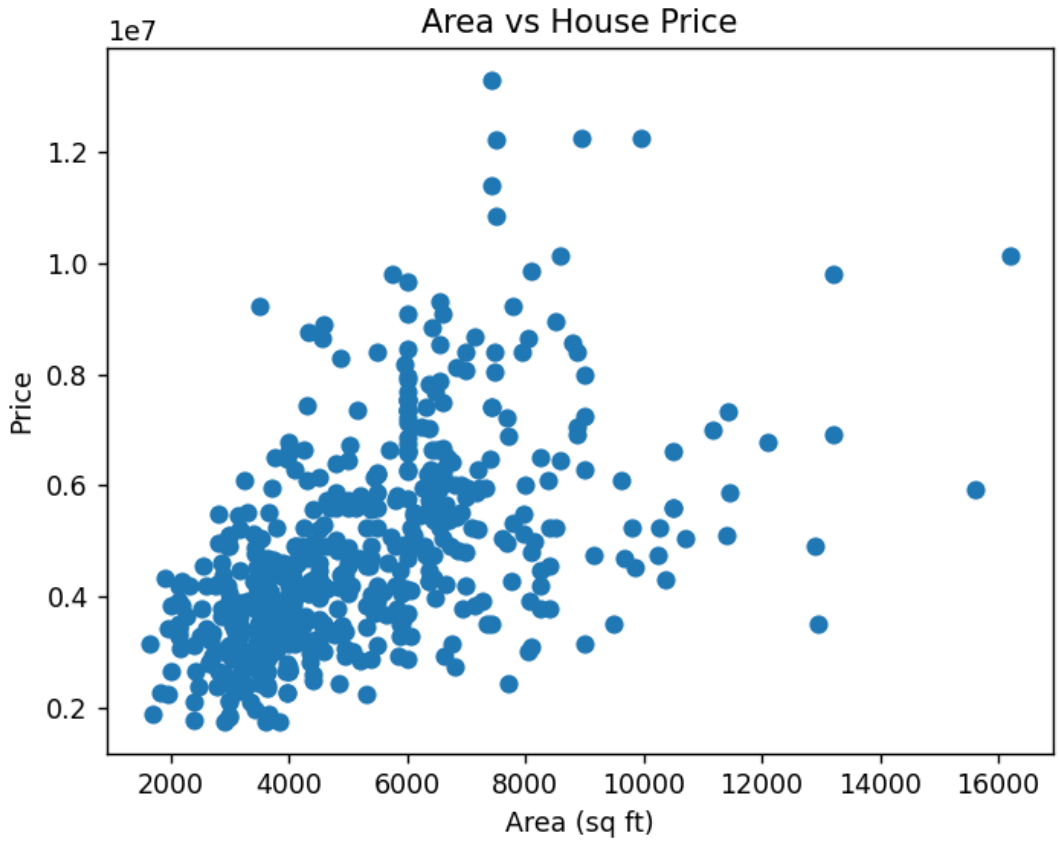


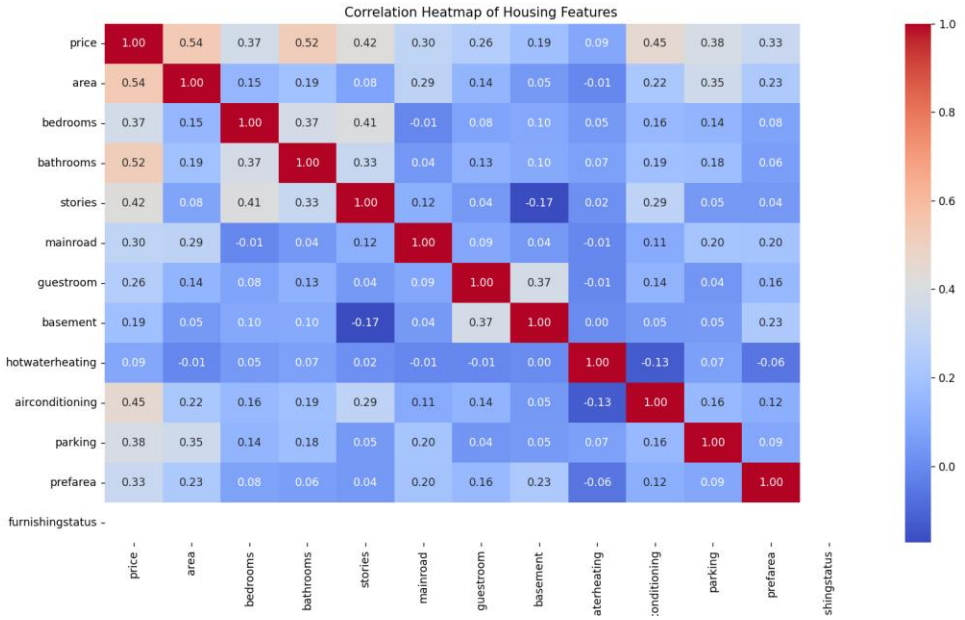
Scenario – 2:


Distribution of Glucose Levels
Frequency
Glucose Level
0
20
40
60
80
40
60
80
100
120
140
160
180
200

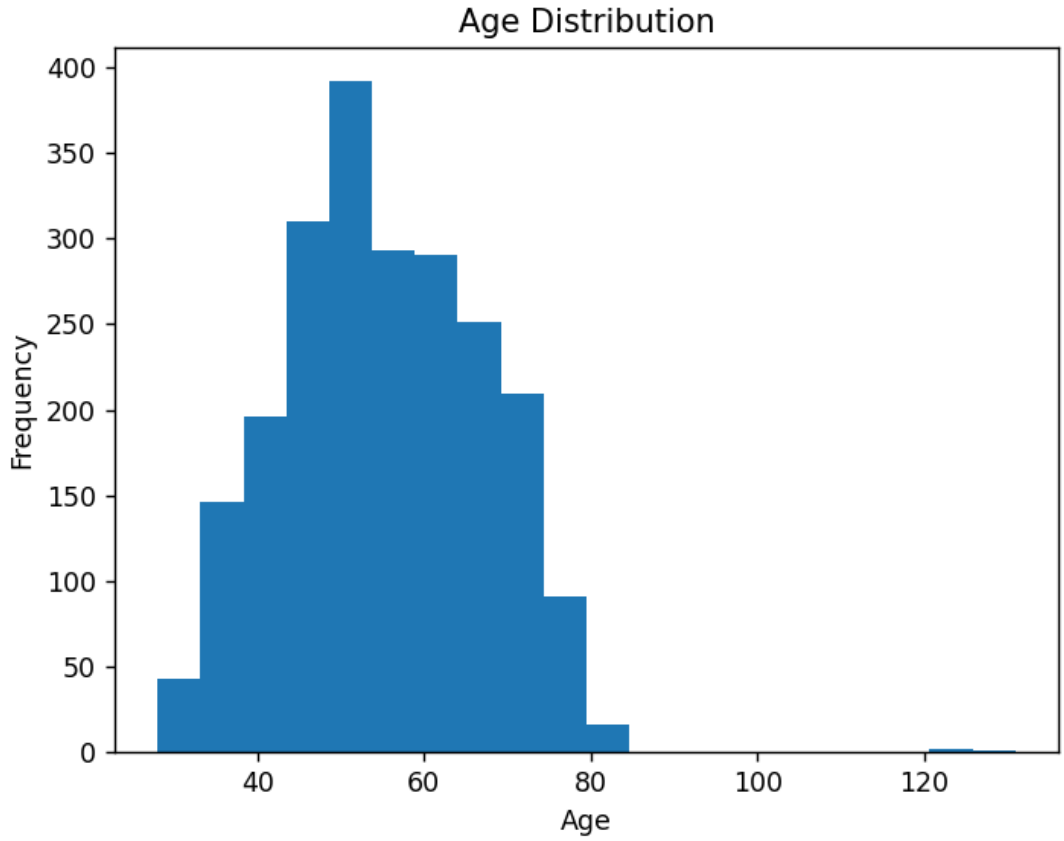

Age Distribution of Patients
Age
20
30
40
50
60
70
80
1

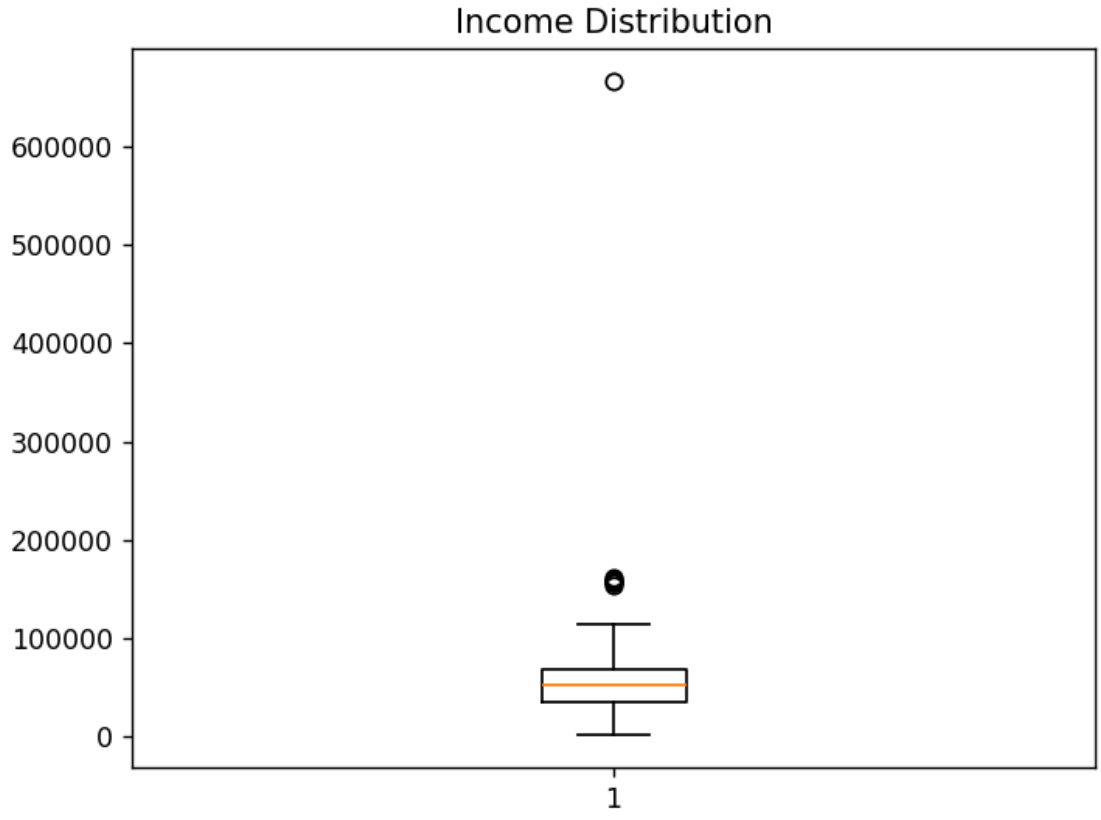
Scenario – 3:


Area vs House Price
1e7
Price
Area (sq ft)
0.2
0.4
0.6
0.8
1.0
1.2
2000
4000
6000
8000
10000
12000
14000
16000


Correlation Heatmap of Housing Features
price 1.00 0.54 0.37 0.52 0.42 0.30 0.26 0.19 0.09 0.45 0.38 0.33
area 0.54 1.00 0.15 0.19 0.08 0.29 0.14 0.05 -0.01 0.22 0.35 0.23
bedrooms 0.37 0.15 1.00 0.37 0.41 -0.01 0.08 0.10 0.05 0.16 0.14 0.08
bathrooms 0.52 0.19 0.37 1.00 0.33 0.04 0.13 0.10 0.07 0.19 0.18 0.06
stories 0.42 0.08 0.41 0.33 1.00 0.12 0.04 -0.17 0.02 0.29 0.05 0.04
mainroad 0.30 0.29 -0.01 0.04 0.12 1.00 0.09 0.04 -0.01 0.11 0.20 0.20
guestroom 0.26 0.14 0.08 0.13 0.04 0.09 1.00 0.37 -0.01 0.14 0.04 0.16
basement 0.19 0.05 0.10 0.10 -0.17 0.04 0.37 1.00 0.00 0.05 0.05 0.23
hotwaterheating 0.09 -0.01 0.05 0.07 0.02 -0.01 -0.01 0.00 1.00 -0.13 0.07 -0.06
airconditioning 0.45 0.22 0.16 0.19 0.29 0.11 0.14 0.05 -0.13 1.00 0.16 0.12
parking 0.38 0.35 0.14 0.18 0.05 0.20 0.04 0.05 0.07 0.16 1.00 0.09
prefarea 0.33 0.23 0.08 0.06 0.04 0.20 0.16 0.23 -0.06 0.12 0.09 1.00
furnishingstatus
price
area
bedrooms
bathrooms
stories
mainroad
guestroom
basement
aterheating
:onditioning
parking
prefarea
shingstatus
1.0
0.8
0.6
0.4
0.2
0.0

Scenario – 4:


Age Distribution
Frequency
Age
0
50
100
150
200
250
300
350
400
40
60
80
100
120


Income Distribution
0
100000
200000
300000
400000
500000
600000
1