**ASSIGNMENT NO. 4**

**AIM :** **Case Study: Visualizing Real Estate Data Insights.**

**This case study focuses creating and interpreting visualizations: histograms for feature distributions, scatter plots to reveal relationships between features and house prices, and heatmaps to visualize correlation matrices.**

The dataset contains 7 columns.

import pandas as pd

# Load the dataset

real\_estate\_data = pd.read\_csv("/Real\_Estate.csv")

# Display the first few rows of the dataset and the info about the dataset

real\_estate\_data\_head = real\_estate\_data.head(6)

data\_info = real\_estate\_data.info()

print(real\_estate\_data\_head)

print(data\_info)

1. **Transaction date**: The date of the real estate transaction.
2. **House age**: Age of the house in years.
3. **Distance to the nearest MRT station**: Distance to the nearest Mass Rapid Transit station in meters.
4. **Number of convenience stores**: Number of convenience stores in the vicinity.
5. **Latitude**: Latitude of the property location.
6. **Longitude**: Longitude of the property location.
7. **House price of unit area**: House price of unit area.

Now, let’s have a look if the data contains any null values or not:

print(real\_estate\_data.isnull().sum())

There are no null values in the dataset.

Now, let’s have a look at the histograms of all the numerical features:

import matplotlib.pyplot as plt

import seaborn as sns

# Set the aesthetic style of the plots

sns.set\_style("whitegrid")

# Create histograms for the numerical columns

fig, axes = plt.subplots(nrows=3, ncols=2, figsize=(12, 12))

fig.suptitle('Histograms of Real Estate Data', fontsize=16)

cols = ['House age', 'Distance to the nearest MRT station', 'Number of convenience stores',

        'Latitude', 'Longitude', 'House price of unit area']

for i, col in enumerate(cols):

    sns.histplot(real\_estate\_data[col], kde=True, ax=axes[i//2, i%2])

    axes[i//2, i%2].set\_title(col)

    axes[i//2, i%2].set\_xlabel('')

    axes[i//2, i%2].set\_ylabel('')

plt.tight\_layout(rect=[0, 0.03, 1, 0.95])

plt.show()

The histograms provide insights into the distribution of each variable:

1. **House Age**: This shows a relatively uniform distribution with a slight increase in the number of newer properties (lower age).
2. **Distance to the Nearest MRT Station**: Most properties are located close to an MRT station, as indicated by the high frequency of lower distances. There’s a long tail extending towards higher distances, suggesting some properties are quite far from MRT stations.
3. **Number of Convenience Stores**: Displays a wide range, with notable peaks at specific counts, like 0, 5, and 10. It suggests certain common configurations in terms of convenience store availability.
4. **Latitude and Longitude**: Both show relatively concentrated distributions, indicating that the properties are located in a geographically limited area.
5. **House Price of Unit Area**: Displays a right-skewed distribution, with a concentration of properties in the lower price range and fewer properties as prices increase.

Creating scatter plots to explore the relationships between these variables and the house price. It will help us understand which factors might be influencing property prices more significantly:

# Scatter plots to observe the relationship with house price

fig, axes = plt.subplots(nrows=2, ncols=2, figsize=(12, 10))

fig.suptitle('Scatter Plots with House Price of Unit Area', fontsize=16)

# Scatter plot for each variable against the house price

sns.scatterplot(data=real\_estate\_data, x='House age', y='House price of unit area', ax=axes[0, 0])

sns.scatterplot(data=real\_estate\_data, x='Distance to the nearest MRT station', y='House price of unit area', ax=axes[0, 1])

sns.scatterplot(data=real\_estate\_data, x='Number of convenience stores', y='House price of unit area', ax=axes[1, 0])

sns.scatterplot(data=real\_estate\_data, x='Latitude', y='House price of unit area', ax=axes[1, 1])

plt.tight\_layout(rect=[0, 0.03, 1, 0.95])

plt.show()

The scatter plots revealed interesting relationships between various factors and house prices:

1. **House Age vs. House Price**: There doesn’t seem to be a strong linear relationship between house age and price. However, it appears that very new and very old houses might have higher prices.
2. **Distance to the Nearest MRT Station vs. House Price**: There is a clear trend showing that as the distance to the nearest MRT station increases, the house price tends to decrease. It suggests a strong negative relationship between these two variables.
3. **Number of Convenience Stores vs. House Price**: There seems to be a positive relationship between the number of convenience stores and house prices. Houses with more convenience stores in the vicinity tend to have higher prices.
4. **Latitude vs. House Price**: While not a strong linear relationship, there seems to be a pattern where certain latitudes correspond to higher or lower house prices. It could be indicative of specific neighbourhoods being more desirable.

Finally, it would be beneficial to perform a correlation analysis to quantify the relationships between these variables, especially how each one correlates with the house price:

# Correlation matrix

# Excluding the 'Transaction date' column which is of type object (string)

correlation\_matrix = real\_estate\_data.select\_dtypes(include=['number']).corr()

# Plotting the correlation matrix

plt.figure(figsize=(10, 6))

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=.5)

plt.title('Correlation Matrix')

plt.show()

print(correlation\_matrix)

The correlation matrix provides quantified insights into how each variable is related to the others, especially with respect to the house price:

1. **House Age**: This shows a very weak negative correlation with house price (-0.012), implying that age is not a strong predictor of price in this dataset.
2. **Distance to Nearest MRT Station**: Has a strong negative correlation with house price (-0.637). It indicates that properties closer to MRT stations tend to have higher prices, which is a significant factor in property valuation.
3. **Number of Convenience Stores**: Displays a moderate positive correlation with house price (0.281). More convenience stores in the vicinity seem to positively affect property prices.
4. **Latitude and Longitude**: Both show a weak correlation with house prices. Latitude has a slight positive correlation (0.081), while longitude has a slight negative correlation (-0.099).

**CONCLUSION :** Overall, the most significant factors affecting house prices in this dataset appear to be the proximity to MRT stations and the number of convenience stores nearby. The geographical location (latitude and longitude) and the age of the house seem to have less impact on the price.