**Question 1:**

**Code:**

import pandas as pd

import matplotlib.pyplot as plt

# Load the dataset (replace 'path\_to\_dataset' with the actual path to your CSV file)

housing\_data = pd.read\_csv('/content/drive/MyDrive/Colab Notebooks/housing.csv')

# Display the first few rows of the dataset to understand its structure

print(housing\_data.head())

# Question 1: What is the distribution of housing prices?

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

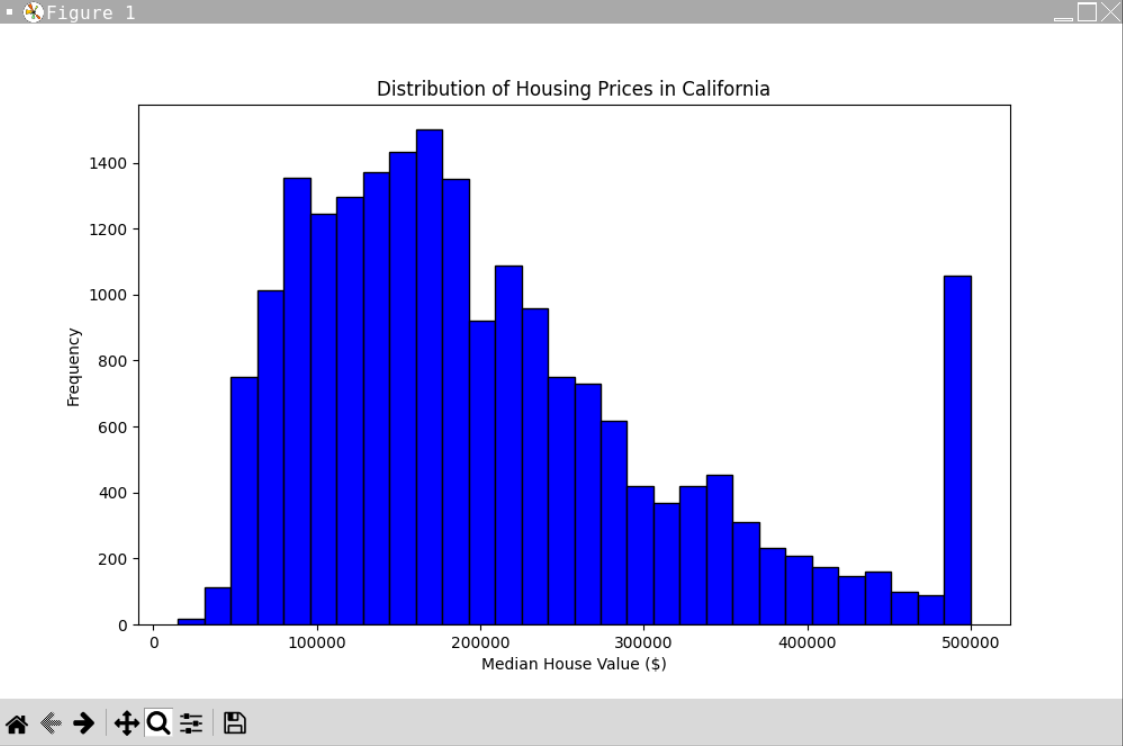
plt.hist(housing\_data['median\_house\_value'], bins=30, color='blue', edgecolor='black')

plt.title('Distribution of Housing Prices in California')

plt.xlabel('Median House Value ($)')

plt.ylabel('Frequency')

plt.show()



**Analysis for Question 1:**

The histogram depicts the dispersion of California house prices. There appear to be a substantial number of houses with a median value in a certain range, with fewer houses at higher and lower values. The bin indicating the most common range of median house values has the most frequency.

The distribution of home prices in California is shown by the histogram, where the y-axis indicates frequency and the x-axis shows median house values in dollars. With a peak frequency for homes in the $100,000–$300,000 range, the majority of the houses are grouped in the lower to moderate price range. There are relatively few homes in the highest price range displayed, which is above $500,000. Prices drop down as they rise. The distribution is right-skewed, meaning that there is a long tail of expensive homes despite the fact that most of the housing stock is less expensive. There is a prominent outlier on the far right, which may indicate that some homes are priced far higher than others. The histogram indicates a broad range of home prices, with the lower to intermediate price ranges being the most affordable.

**Question 2: How does the median income relate to the median house value?**

**Code:**

import pandas as pd

import matplotlib.pyplot as plt

# Load the dataset

housing\_data = pd.read\_csv('housing.csv')

# Display the first few rows of the dataset to understand its structure

print(housing\_data.head())

# Question 2: How does the median income relate to the median house value?

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

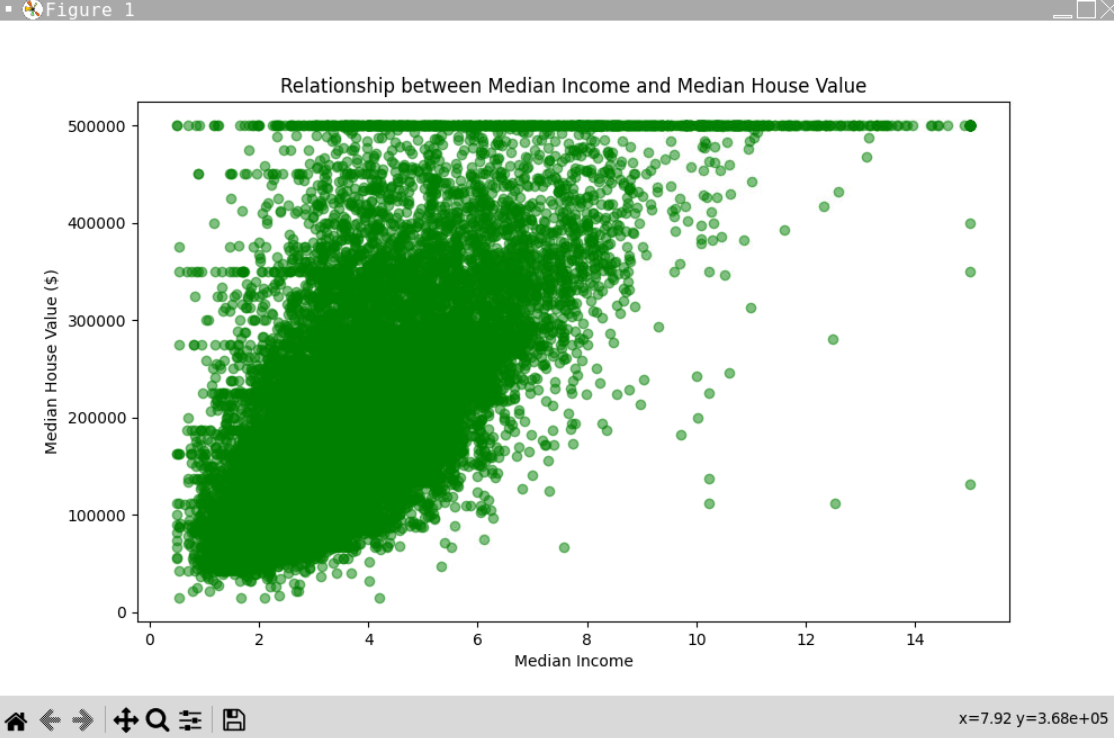
plt.scatter(housing\_data['median\_income'], housing\_data['median\_house\_value'], alpha=0.5, color='green')

plt.title('Relationship between Median Income and Median House Value')

plt.xlabel('Median Income')

plt.ylabel('Median House Value ($)')

plt.show()



Analysis: The correlation between median income and median house value is depicted in the scatter plot. Given that higher income is typically correlated with higher housing values, this indicates a positive association between income and value. At the lower end of the income spectrum, the data points are closely spaced, indicating a high number of observations with lower median incomes and home values. The gap of home values widens with income, suggesting greater volatility in home prices at higher income levels. The lack of a distinct upper limit for home values at higher income levels may indicate that the high-end housing market is influenced by variables other than median income. Although there isn't a clear linear link between the two, the plot does indicate that median income is a strong predictor of median housing value.

**Question 3: Explore patterns or trends in the data through exploratory data analysis and visualization.**

**Code:**

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

# Loading the dataset

housing\_data = pd.read\_csv('/content/drive/MyDrive/Colab Notebooks/housing.csv')

# Display the first few rows of the dataset to understand its structure

print(housing\_data.head())

# Question 3: Explore patterns or trends in the data through exploratory data analysis and visualization

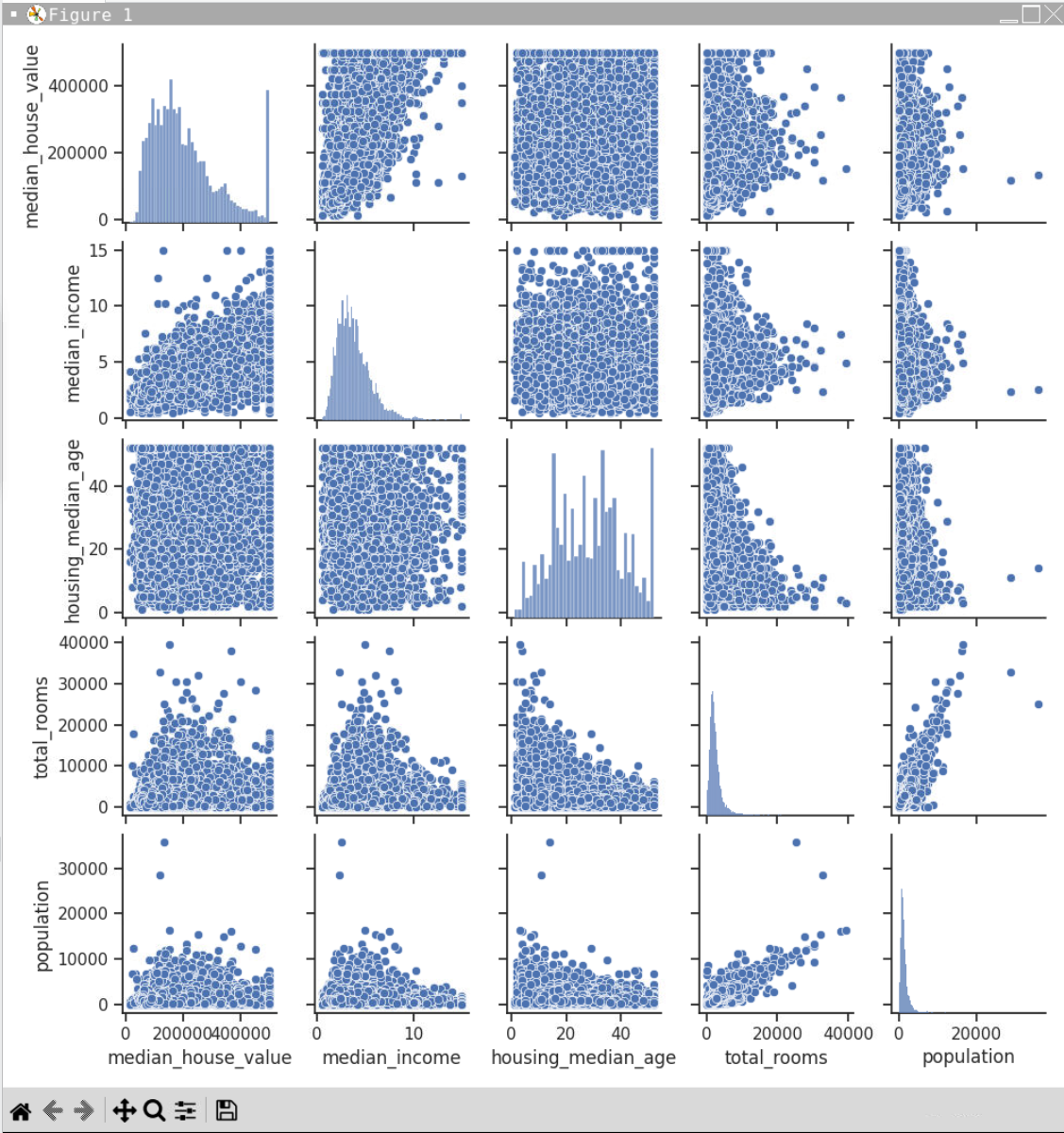
# Create a pair plot to visualize relationships between multiple variables

sns.set(style="ticks")

sns.pairplot(housing\_data[['median\_house\_value', 'median\_income', 'housing\_median\_age', 'total\_rooms', 'population']], height=2)

plt.suptitle("Pair Plot of Selected Variables", y=1.02)

plt.show()



**Analysis:**

A thorough summary of the correlations between the variables in a dataset linked to housing can be obtained from the pair plot. Remarkably, the distributions of median income and median house value are both right-skewed, suggesting that lower-income and lower-priced residences occur more frequently. The more consistent distribution of housing median age, on the other hand, points to a wide variety of housing ages. The distributions of total rooms and population are also right-skew, indicating a predominance of smaller homes and sparserly populated places.The idea that higher income levels typically correspond with more expensive housing is supported by the bivariate scatter plots, which show a strong positive association between median income and median house value. One important lesson from this relationship is that economic status plays a significant role in house affordability. On the other hand, the correlations between population and total rooms and median house value are more nuanced, suggesting a complicated web of interrelated factors influencing housing costs. There are observable trends and anomalies in these charts. Larger houses are generally more expensive, although there is a lot of variation in this relationship, as evidenced by the weakly positive association found between the number of rooms and the median house value. There are clear outliers, such as regions with unusually high population densities compared to room counts, which may indicate densely crowded housing.The matrix emphasises how important median income is in determining the value of a home, but it also shows how a variety of characteristics, including size, age, and population density, have an impact on housing economics. The pair plot proves to be a highly useful instrument for analysing the dynamics of the housing market when these variables are examined together. This highlights the significance of multifactorial analysis in comprehending sectoral trends.

**Question 4: What is the geographical distribution of high-value houses?**

**Code:**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

housing\_data = pd.read\_csv('/content/drive/MyDrive/Colab Notebooks/housing.csv')

# Display the first few rows of the dataset to understand its structure

print(housing\_data.head())

# Question 4: What is the geographical distribution of high-value houses?

plt.figure(figsize=(12, 8))

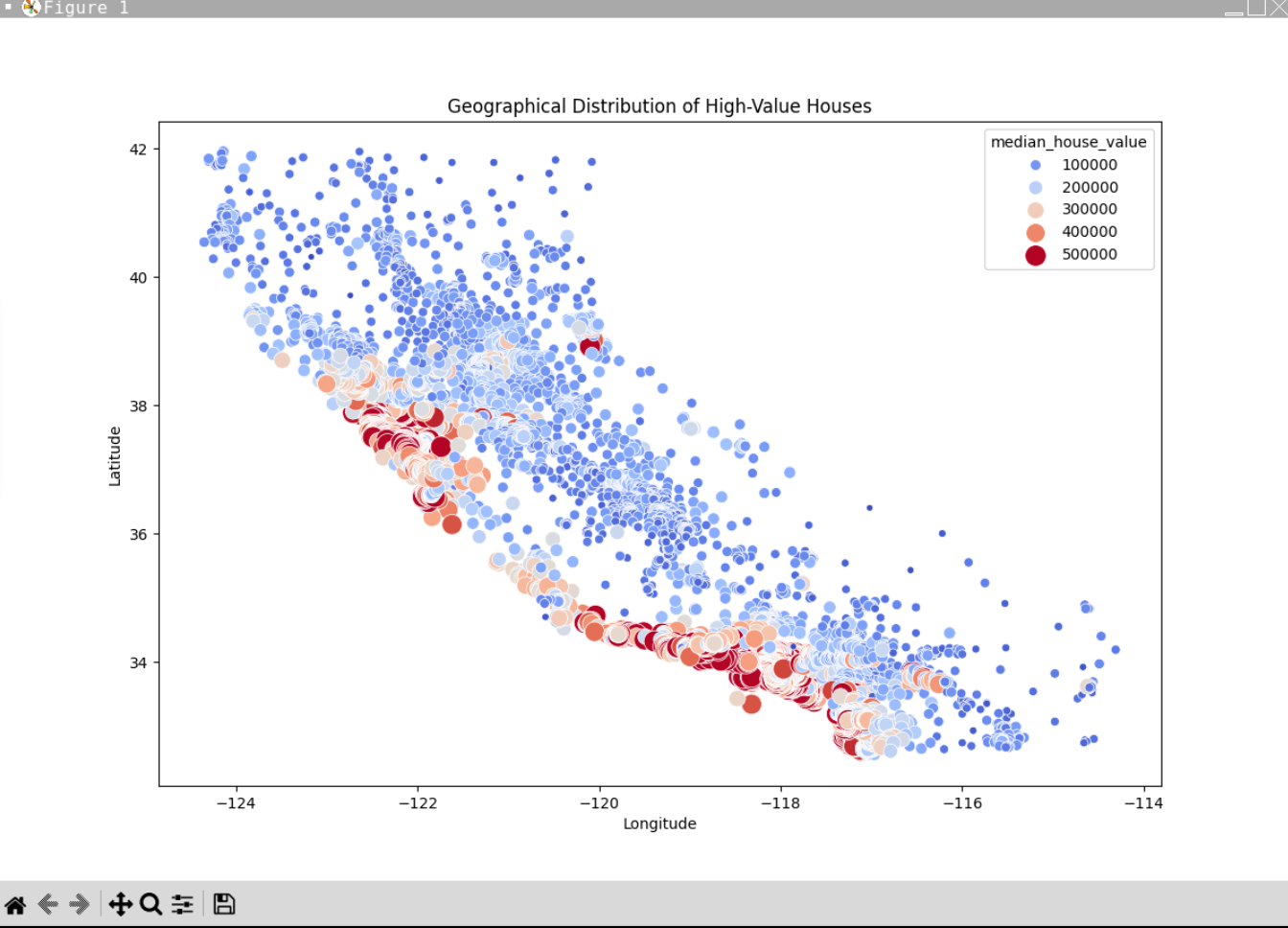
sns.scatterplot(x='longitude', y='latitude', hue='median\_house\_value', palette='coolwarm', size='median\_house\_value', sizes=(20, 200), data=housing\_data)

plt.title('Geographical Distribution of High-Value Houses')

plt.xlabel('Longitude')

plt.ylabel('Latitude')

plt.show()



**Analysis:**

Based on the given latitude and longitude coordinates, this map shows the geographic distribution of house values in an area that mimics California. Different colours denote different price ranges, and each dot reflects the area's median house value. Dark red dots indicate homes valued about $500,000, while light blue dots indicate dwellings of $100,000 in value.

Upon examining the map, we can observe a large number of light blue dots dispersed over the area, indicating that $100,000 residences are prevalent and dispersed. The dots on the map show that the value of a home increases from light blue to orange to dark red. These higher-priced residences appear to be concentrated along the beach, especially in the southern region. Because of the allure of living close to the water, this clustering implies that more expensive properties are clustered in coastal locations, which is frequently the case in real life.

The chart also demonstrates how fewer homes are available and less dispersed around the area as property values rise. Rather, they congregate at specific areas. These premium locations are probably well-known cities or neighbourhoods that appeal due to their prominence, facilities, or convenient location.

**Question 5: Analyze the distribution of total rooms and total bedrooms in the dataset.**

**Code:**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

housing\_data = pd.read\_csv('/content/drive/MyDrive/Colab Notebooks/housing.csv')

# Display the first few rows of the dataset to understand its structure

print(housing\_data.head())

# Question 5: Analyze the distribution of total rooms and total bedrooms in the dataset.

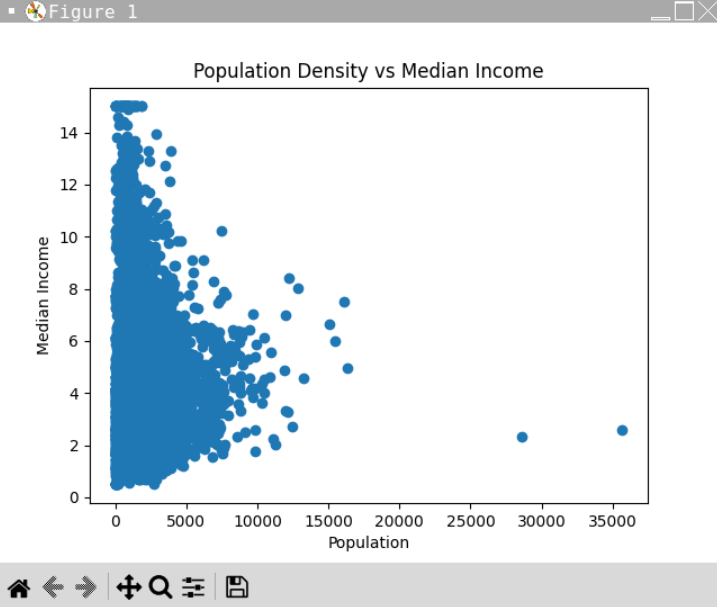
plt.scatter(housing\_data['population'], housing\_data['median\_income'])

plt.xlabel('Population')

plt.ylabel('Median Income')

plt.title('Population Density vs Median Income')

plt.show()



**Analysis:**

This scatter plot displays two variables: the average income of the residents of a given area and the population. The number of persons is indicated by the bottom line of the chart, which runs from left to right. The average income is shown on the chart's rising side.

The majority of the dots are on the left side, towards the beginning of the bottom line. This indicates that most localities are not very populous and that the income distribution among them is highly variable, with some making relatively little and others very much. The dots don't rise as high when we turn to look to the right, where there are areas with more people. This indicates that in densely populated areas, the average income is usually not excessive.

A few dots with plenty of people and low average income are located far to the right. Even if people make money, it could not be much after they pay for their housing and other expenses in these large, expensive cities.

It is not evident from the chart's shape that an increase or decrease in the number of persons corresponds to a change in income. Rather, it demonstrates how varied places are. Places with a large population tend to be more middle-class financially, while areas with few residents may be extremely wealthy or extremely impoverished.

This image illustrates that there is no hard-and-fast rule that says having more people equals having more money. It's more intricate, and there must be other factors influencing people's income.

**Question 6: Can we identify any outliers in the dataset, and how do they impact our analysis?**

**Code:**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

housing\_data = pd.read\_csv('/content/drive/MyDrive/Colab Notebooks/housing.csv')

# Calculate Z-scores for 'median\_house\_value'

z\_scores = (housing\_data['median\_house\_value'] - housing\_data['median\_house\_value'].mean()) / housing\_data['median\_house\_value'].std()

# Set a threshold for Z-scores to identify outliers (e.g., Z-score > 3 or < -3)

outliers\_threshold = 3

outliers = housing\_data[abs(z\_scores) > outliers\_threshold]

# Visualize outliers using a box plot

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

sns.boxplot(x='median\_house\_value', data=housing\_data)

plt.title('Box Plot of Median House Value with Outliers Highlighted')

plt.scatter(outliers.index, outliers['median\_house\_value'], color='r', label='Outliers')

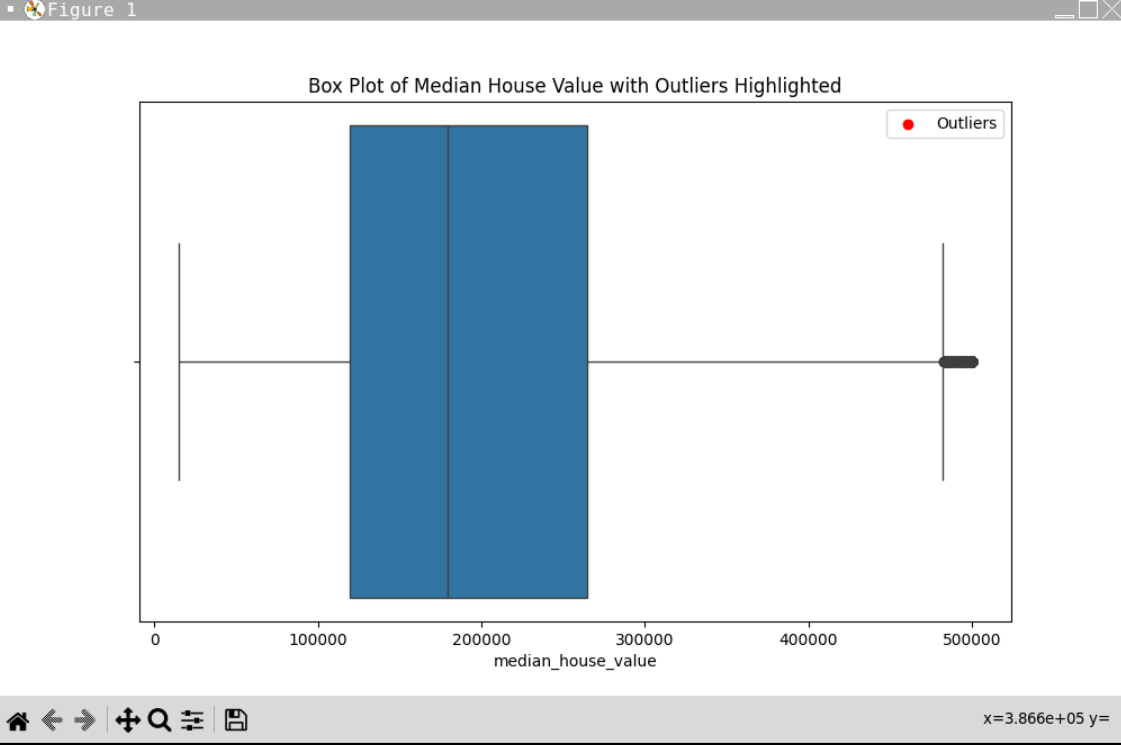
plt.legend()

plt.show()

# Display the outliers

print("Outliers:")

print(outliers)



**Analysis:**

A straightforward method of estimating the cost of a house in a location is to use the graph with the blue box. The more costly homes are located at the top of the box, and the lesser ones are located at the bottom. The average price is indicated by the line in the centre of the box. The houses with the longest lines that protrude from the box are the most expensive and least expensive, but there aren't many that are. Additionally, a unique red dot represents a house that is far more expensive than the others and so unique that it belongs in a different category. Without having to examine every single one, this blue box image aids in our understanding of housing values.

**Query 1: Query to plot the distribution of housing prices.**

**Code:**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

housing\_data = pd.read\_csv('/content/drive/MyDrive/Colab Notebooks/housing.csv')

# Query to plot the distribution of housing prices.

plt.figure(figsize=(8, 5))

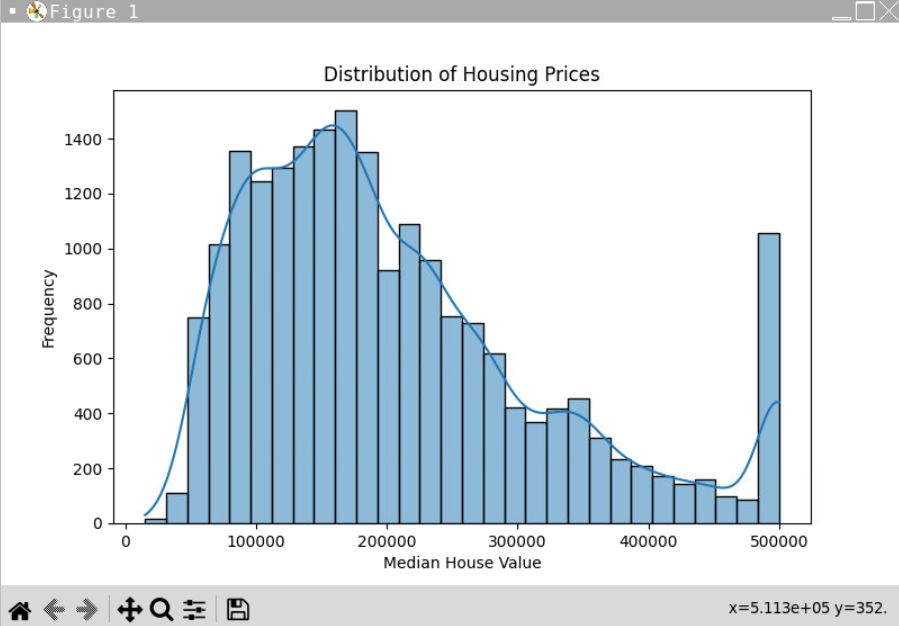
sns.histplot(housing\_data['median\_house\_value'], bins=30, kde=True)

plt.title('Distribution of Housing Prices')

plt.xlabel('Median House Value')

plt.ylabel('Frequency')

plt.show()



Analysis:

This histogram, looks like it's showing the distribution of housing prices. On the x-axis (the horizontal line at the bottom), we have 'Median House Value,' which means the middle value of house prices in certain categories. The y-axis (the vertical line on the left) shows 'Frequency,' which tells us how many houses fall into each price category. The bars go up in height according to how many houses are in each price range. So, a taller bar means more houses at that price. There's also a smooth curve over the bars, which helps to show the overall shape of the distribution. It looks like most of the houses are in the lower to middle price ranges, with fewer houses in the very cheap or very expensive categories. There's a note on the chart with an 'x' and 'y' value, probably where the mouse was pointing when the picture was taken. This kind of chart helps people understand what prices are most common when they're looking to buy a house, for example.

**Query 2: Query for calculating the correlation between median income and median house value.**

**Code:**

import pandas as pd

import matplotlib.pyplot as plt

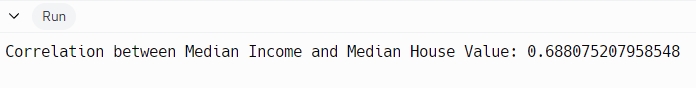
import seaborn as sns

housing\_data = pd.read\_csv('/content/drive/MyDrive/Colab Notebooks/housing.csv')

# query for calculating the correlation between median income and median house value

correlation = housing\_data['median\_income'].corr(housing\_data['median\_house\_value'])

print(f'Correlation between Median Income and Median House Value: {correlation}')



**Analysis:**

The correlation coefficient of approximately 0.69, indicating a moderate positive relationship between median income and median house value.

**Query 3: Query for creating scatter plots for housing characteristics vs. prices**

**Code:**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

housing\_data = pd.read\_csv('/content/drive/MyDrive/Colab Notebooks/housing.csv')

# Query for creating scatter plots for housing characteristics vs. prices

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

axes[0, 0].scatter(housing\_data['housing\_median\_age'], housing\_data['median\_house\_value'])

axes[0, 0].set\_xlabel('Housing Median Age')

axes[0, 0].set\_ylabel('Median House Value')

axes[0, 1].scatter(housing\_data['total\_rooms'], housing\_data['median\_house\_value'])

axes[0, 1].set\_xlabel('Total Rooms')

axes[0, 1].set\_ylabel('Median House Value')

axes[1, 0].scatter(housing\_data['population'], housing\_data['median\_house\_value'])

axes[1, 0].set\_xlabel('Population')

axes[1, 0].set\_ylabel('Median House Value')

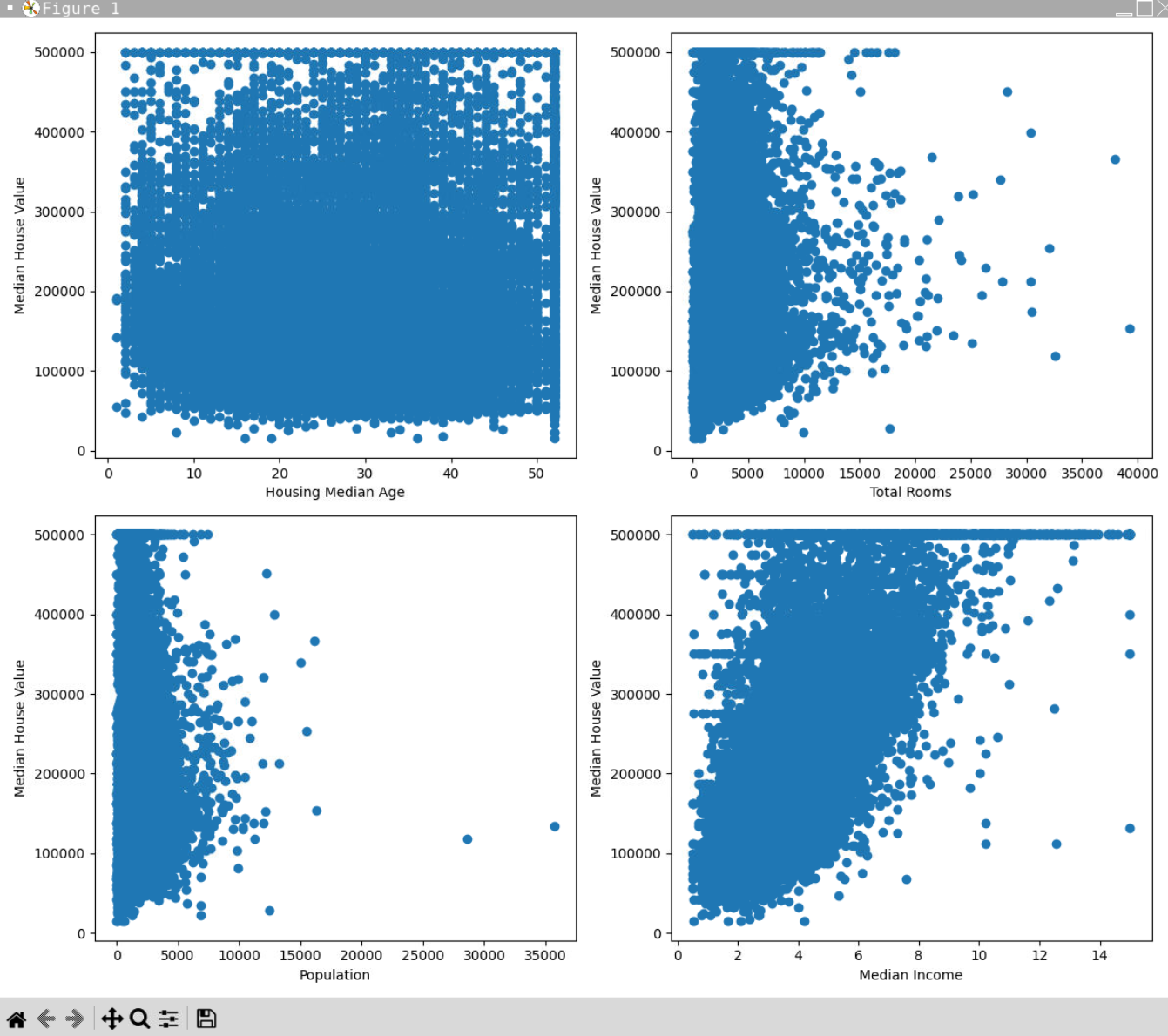
axes[1, 1].scatter(housing\_data['median\_income'], housing\_data['median\_house\_value'])

axes[1, 1].set\_xlabel('Median Income')

axes[1, 1].set\_ylabel('Median House Value')

plt.tight\_layout()

plt.show()



**Analysis:**

The four scatter plots shows, which are graphs that use dots to show how two sets of data are related to each other. Each plot has lots of dots, and each dot represents a different area or neighborhood. The first plot on the top left doesn't show much of a pattern; it's about how old houses are and their prices, but the dots are all over the place. The top right plot shows that as the number of rooms in houses goes up, so does the price, with the dots going up in a clear direction. The bottom left plot is about how many people live in an area and the price of houses. Here, when there are more people, it doesn't really mean higher prices, as the dots are spread out widely. The last plot on the bottom right is about how much money people make and the price of houses. This one shows that as people earn more money, the price of houses tends to be higher, with the dots forming a band that goes up to the right. At the bottom of this plot, there's some text that probably shows where the mouse was pointing when the picture was taken. Scatter plots like these help us see if there's a link between things, like if having more rooms or making more money might mean more expensive houses.