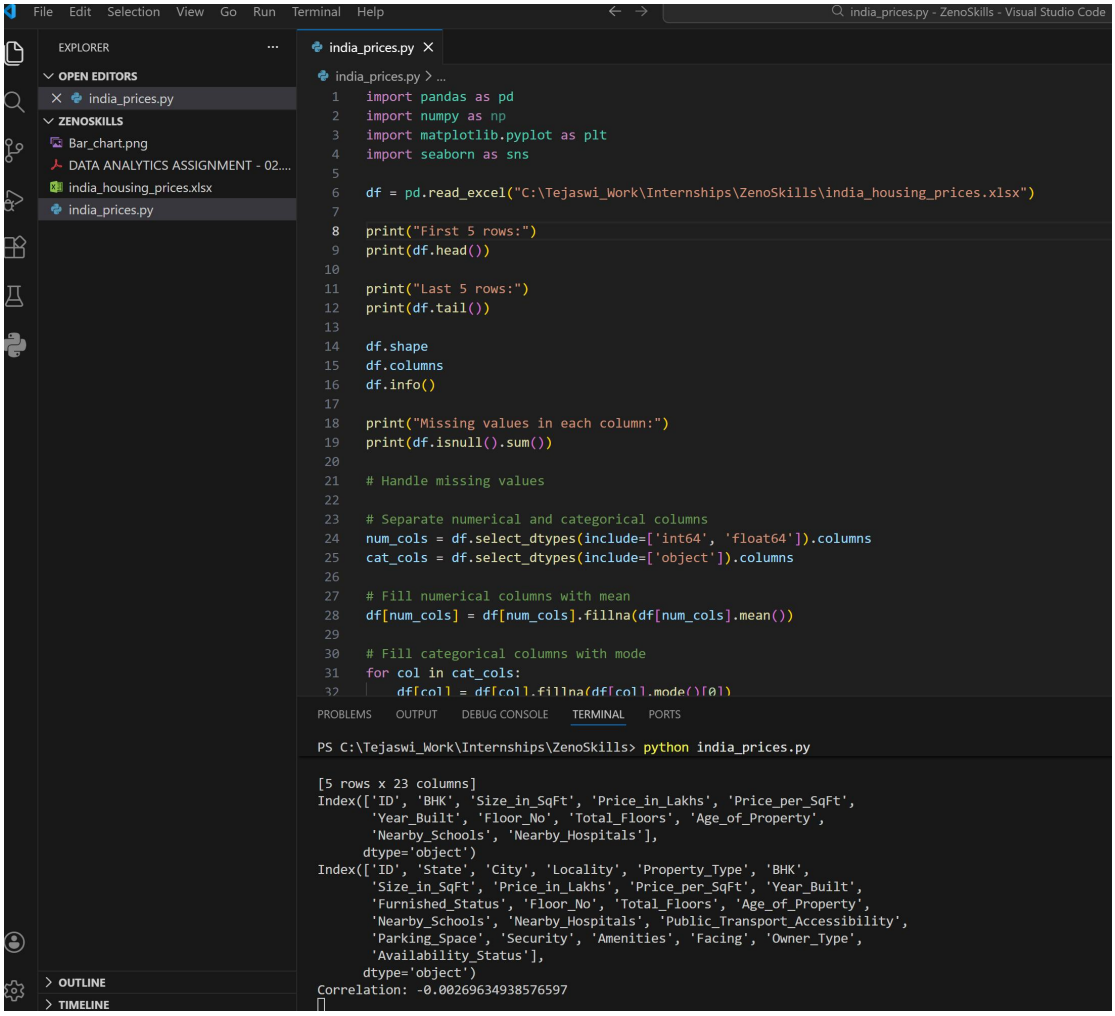
Python File

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_excel("C:\Tejaswi_Work\Internships\ZenoSkills\india_housing_prices.xlsx")



print("First 5 rows:")
print(df.head())

print("Last 5 rows:")
print(df.tail())

df.shape
df.columns
df.info()

print("Missing values in each column:")
print(df.isnull().sum())

# Handle missing values

# Separate numerical and categorical columns
num_cols = df.select_dtypes(include=['int64', 'float64']).columns
cat_cols = df.select_dtypes(include=['object']).columns

# Fill numerical columns with mean
df[num_cols] = df[num_cols].fillna(df[num_cols].mean())

# Fill categorical columns with mode
for col in cat_cols:
    df[col] = df[col].fillna(df[col].mode()[0])

 #3 Check and remove duplicate records

print("\nDuplicate rows count:", df.duplicated().sum())

# Convert numerical columns explicitly (example)
# Replace with your column names if needed
#df['price'] = df['price'].astype(float)
#df['size_in_sqft'] = df['size_in_sqft'].astype(int)


#Standardize text data
```

```
for col in cat_cols:
    df[col] = df[col].str.strip().str.title()



# Summary statistics for numerical columns
print("Summary statistics (Numerical):")
print(df.describe())

# Summary statistics for categorical columns
print("\nSummary statistics (Categorical):")
print(df.describe(include='object'))


cat_cols = df.select_dtypes(include='object').columns
# Value counts for each categorical column
for col in cat_cols:
    print(f"\nValue counts for {col}:")
    print(df[col].value_counts())

df['City']

df.groupby('City')['Price_in_Lakhs'].mean()
df.groupby('City')['Price_in_Lakhs'].sum()
df.groupby('City')['Price_in_Lakhs'].count()

group_summary = df.groupby('City')['Price_in_Lakhs'].agg(
    Average_Price='mean',
    Total_Price='sum',
    Property_Count='count'
)

print(group_summary)


print(" top 5 categories:")
print(df.head())

print(" bottom 5 categories:")
print(df.head())



# Select only numerical columns
num_cols = df.select_dtypes(include=['int64', 'float64'])

print(num_cols.columns)
print(df.columns)


correlation = df['Nearby_Schools'].corr(df['Price_in_Lakhs'])
print("Correlation:", correlation)

#Bar chart
city_avg_price = df.groupby('City')['Price_in_Lakhs'].mean()
```

```
plt.figure()
city_avg_price.plot(kind='bar')
plt.xlabel("City")
plt.ylabel("Average Price")
plt.title("Average Price by City")
plt.show()


#Line chart
year_price = df.groupby('Year_Built')['Price_in_Lakhs'].mean()
plt.figure()
plt.plot(year_price.index, year_price.values)
plt.xlabel("Year")
plt.ylabel("Average Price")
plt.title("Price Trend Over Years")
plt.show()

#Histogram
plt.figure()
plt.hist(df['Price_in_Lakhs'], bins=20)
plt.xlabel("Price")
plt.ylabel("Frequency")
plt.title("Price Distribution")
plt.show()

#ScotterPlot
plt.figure()
plt.scatter(df['Property_Type'], df['Price_in_Lakhs'])
plt.xlabel("Property_Type")
plt.ylabel("Price")
plt.title("Size vs Price")
plt.show()

#Box Plot
plt.figure()
plt.boxplot(df['Floor_No'])
plt.ylabel("Floor_No")
plt.title("Box Plot of Price")
plt.show()

# Select numerical columns
num_df = df.select_dtypes(include=['int64', 'float64'])

# Correlation matrix
corr = num_df.corr()

plt.figure()
plt.imshow(corr)
plt.colorbar()
plt.xticks(range(len(corr.columns)), corr.columns, rotation=90)
plt.yticks(range(len(corr.columns)), corr.columns)
plt.title("Correlation Heatmap")
plt.show()

#Experiment
sns.heatmap(df[['Nearby_Schools', 'Price_in_Lakhs']].corr(), annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Correlation: Nearby Schools vs Price")
```

plt.show()


File Edit Selection View Go Run Terminal Help
india_prices.py - ZenoSkills - Visual Studio Code
EXPLORER
OPEN EDITORS
india_prices.py
ZENOSKILLS
Bar_chart.png
DATA ANALYTICS ASSIGNMENT - 02....
india_housing_prices.xlsx
india_prices.py
OUTLINE
TIMELINE
india_prices.py
india_prices.py > ...
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import seaborn as sns
5
6 df = pd.read_excel("C:\Tejaswi_Work\Internships\ZenoSkills\india_housing_prices.xlsx")
7
8 print("First 5 rows:")
9 print(df.head())
10
11 print("Last 5 rows:")
12 print(df.tail())
13
14 df.shape
15 df.columns
16 df.info()
17
18 print("Missing values in each column:")
19 print(df.isnull().sum())
20
21 # Handle missing values
22
23 # Separate numerical and categorical columns
24 num_cols = df.select_dtypes(include=['int64', 'float64']).columns
25 cat_cols = df.select_dtypes(include=['object']).columns
26
27 # Fill numerical columns with mean
28 df[num_cols] = df[num_cols].fillna(df[num_cols].mean())
29
30 # Fill categorical columns with mode
31 for col in cat_cols:
32 df[col] = df[col].fillna(df[col].mode()[0])
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
PS C:\Tejaswi_Work\Internships\ZenoSkills> python india_prices.py
[5 rows x 23 columns]
Index(['ID', 'BHK', 'Size_in_SqFt', 'Price_in_Lakhs', 'Price_per_SqFt',
'Year_Built', 'Floor_No', 'Total_Floors', 'Age_of_Property',
'Nearby_Schools', 'Nearby_Hospitals'],
dtype='object')
Index(['ID', 'State', 'City', 'Locality', 'Property_Type', 'BHK',
'Size_in_SqFt', 'Price_in_Lakhs', 'Price_per_SqFt', 'Year_Built',
'Furnished_Status', 'Floor_No', 'Total_Floors', 'Age_of_Property',
'Nearby_Schools', 'Nearby_Hospitals', 'Public_Transport_Accessibility',
'Parking_Space', 'Security', 'Amenities', 'Facing', 'Owner_Type',
'Availability_Status'],
dtype='object')
Correlation: -0.00269634938576597