

Data cleaning is an essential step in data preprocessing. Here's a step-by-step guide on how to clean the provided dataset:

Step 1: Import necessary libraries and load the data

```
import pandas as pd
import numpy as np

# Load the data
data = pd.read_csv('your_data.csv')
```

Step 2: Handle missing values

```
# Check for missing values
print(data.isnull().sum())

# Replace missing values in 'brokered_by' column with NaN
data['brokered_by'] = data['brokered_by'].replace('-', np.nan)

# Replace missing values in 'price' column with NaN
data['price'] = data['price'].replace(['', '-'], np.nan)

# Replace missing values in 'bed' column with NaN
data['bed'] = data['bed'].replace(['', '-'], np.nan)

# Replace missing values in 'bath' column with NaN
data['bath'] = data['bath'].replace(['', '-'], np.nan)

# Replace missing values in 'acre_lot' column with NaN
data['acre_lot'] = data['acre_lot'].replace(['', '-'], np.nan)

# Replace missing values in 'street' column with NaN
data['street'] = data['street'].replace(['', '-'], np.nan)

# Replace missing values in 'city' column with NaN
data['city'] = data['city'].replace(['', '-'], np.nan)

# Replace missing values in 'state' column with NaN
data['state'] = data['state'].replace(['', '-'], np.nan)

# Replace missing values in 'zip_code' column with NaN
data['zip_code'] = data['zip_code'].replace(['', '-'], np.nan)

# Replace missing values in 'house_size' column with NaN
data['house_size'] = data['house_size'].replace(['', '-'], np.nan)

# Replace missing values in 'prev_sold_date' column with NaN
data['prev_sold_date'] = data['prev_sold_date'].replace(['', '-'], np.nan)
```

Step 3: Remove dollar signs from 'price' column

```
# Remove dollar signs from 'price' column
data['price'] = data['price'].str.replace('$', '').replace('', np.nan).astype(float)
```

Step 4: Convert 'bed', 'bath', and 'acre_lot' columns to numeric

```
# Convert 'bed' column to numeric
data['bed'] = pd.to_numeric(data['bed'], errors='coerce')

# Convert 'bath' column to numeric
data['bath'] = pd.to_numeric(data['bath'], errors='coerce')
```

```
# Convert 'acre_lot' column to numeric
data['acre_lot'] = pd.to_numeric(data['acre_lot'], errors='coerce')
```

Step 5: Convert 'prev_sold_date' column to datetime

```
# Convert 'prev_sold_date' column to datetime
data['prev_sold_date'] = pd.to_datetime(data['prev_sold_date'], errors='coerce')
```

Step 6: Remove rows with missing values in 'status' column

```
# Remove rows with missing values in 'status' column
data = data.dropna(subset=['status'])
```

Step 7: Remove rows with missing values in 'price' column

```
# Remove rows with missing values in 'price' column
data = data.dropna(subset=['price'])
```

Step 8: Save the cleaned data

```
# Save the cleaned data
data.to_csv('cleaned_data.csv', index=False)
```

After running these steps, you should have a cleaned dataset with missing values handled, dollar signs removed from the 'price' column, and columns converted to the correct data types.

Here is the complete code:

```
import pandas as pd
import numpy as np

def clean_data(data):
    # Check for missing values
    print(data.isnull().sum())

    # Replace missing values in 'brokered_by' column with NaN
    data['brokered_by'] = data['brokered_by'].replace('-', np.nan)

    # Replace missing values in 'price' column with NaN
    data['price'] = data['price'].replace(['', '-'], np.nan)

    # Replace missing values in 'bed' column with NaN
    data['bed'] = data['bed'].replace(['', '-'], np.nan)

    # Replace missing values in 'bath' column with NaN
    data['bath'] = data['bath'].replace(['', '-'], np.nan)

    # Replace missing values in 'acre_lot' column with NaN
    data['acre_lot'] = data['acre_lot'].replace(['', '-'], np.nan)

    # Replace missing values in 'street' column with NaN
    data['street'] = data['street'].replace(['', '-'], np.nan)

    # Replace missing values in 'city' column with NaN
    data['city'] = data['city'].replace(['', '-'], np.nan)

    # Replace missing values in 'state' column with NaN
    data['state'] = data['state'].replace(['', '-'], np.nan)

    # Replace missing values in 'zip_code' column with NaN
    data['zip_code'] = data['zip_code'].replace(['', '-'], np.nan)

    # Replace missing values in 'house_size' column with NaN
```

```

data['house_size'] = data['house_size'].replace(['', '-'], np.nan)

# Replace missing values in 'prev_sold_date' column with NaN
data['prev_sold_date'] = data['prev_sold_date'].replace(['', '-'], np.nan)

# Remove dollar signs from 'price' column
data['price'] = data['price'].str.replace('$', '').replace('', np.nan).astype(float)

# Convert 'bed' column to numeric
data['bed'] = pd.to_numeric(data['bed'], errors='coerce')

# Convert 'bath' column to numeric
data['bath'] = pd.to_numeric(data['bath'], errors='coerce')

# Convert 'acre_lot' column to numeric
data['acre_lot'] = pd.to_numeric(data['acre_lot'], errors='coerce')

# Convert 'prev_sold_date' column to datetime
data['prev_sold_date'] = pd.to_datetime(data['prev_sold_date'], errors='coerce')

# Remove rows with missing values in 'status' column
data = data.dropna(subset=['status'])

# Remove rows with missing values in 'price' column
data = data.dropna(subset=['price'])

return data

# Load the data
data = pd.read_csv('your_data.csv')

# Clean the data
cleaned_data = clean_data(data)

# Save the cleaned data
cleaned_data.to_csv('cleaned_data.csv', index=False)

```