

I'll help you clean and prepare this dataset. Here are some data wrangling steps we can perform:

1. Remove rows with all NaN values
2. Handle the bedrooms\_bathrooms column by splitting it into separate columns
3. Convert data types to appropriate formats
4. Clean up any inconsistencies

Here's the Python code to do this:

```
import pandas as pd
import numpy as np

# Assuming your data is in a DataFrame called 'df'
# 1. Remove rows with all NaN values
df = df.dropna(how='all')

# 2. Split bedrooms_bathrooms into separate columns
df[['bedrooms', 'bathrooms']] = df['bedrooms_bathrooms'].str.split(',', expand=True)
df['bedrooms'] = df['bedrooms'].astype(float)
df['bathrooms'] = df['bathrooms'].astype(float)

# 3. Convert data types
df['price'] = df['price'].astype(float)
df['acre_lot'] = df['acre_lot'].astype(float)
df['zip_code'] = df['zip_code'].astype(float)
df['house_size'] = df['house_size'].astype(float)
df['prev_sold_date'] = pd.to_datetime(df['prev_sold_date'])

# 4. Clean up the status column
df['status'] = df['status'].str.lower()

# 5. Create a new column for state from address
df['state'] = df['address'].str.split(',').str[-1].str.strip()

# 6. Create a new column for city from address
df['city'] = df['address'].str.split(',').str[-2].str.strip()

# 7. Calculate price per square foot
df['price_per_sqft'] = df['price'] / df['house_size']

# Drop the original bedrooms_bathrooms column as we've split it
df = df.drop('bedrooms_bathrooms', axis=1)

# Reorder columns in a logical way
columns_order = ['address', 'city', 'state', 'zip_code', 'price', 'status',
                 'bedrooms', 'bathrooms', 'house_size', 'acre_lot',
                 'price_per_sqft', 'prev_sold_date', 'brokered_by']
df = df[columns_order]

# Basic statistics about the cleaned dataset
print("\nDataset Summary:")
print(f"Total number of records: {len(df)}")
print(f"Number of properties for sale: {len(df[df['status'] == 'for_sale'])}")
print(f"Number of sold properties: {len(df[df['status'] == 'sold'])}")
print(f"\nPrice Statistics:")
print(df['price'].describe())
```

This cleaned dataset will now have: - No completely empty rows - Separate columns for bedrooms and bathrooms - Proper data types for numerical and date columns - Additional useful columns like price\_per\_sqft - Separated location information (city and state) - A more organized column structure

Would you like me to perform any additional specific cleaning or analysis on this dataset?